LLNL-AR-411687

SciDAC Earth System Grid Center for Enabling Technologies

Status Report for the Period
October 1, 2006 through March 31, 2009

Principal Investigators
Dean N. Williams[3], Ian T. Foster[1], and Don E. Middleton[5]

The Earth System Grid Center for Enabling Technologies Team:
Rachana Ananthakrishnan[1], Neill Miller[1], Frank Siebenlist[1]
Junmin Gu[2], Vijaya Natarajan[2], Arie Shoshani[2], Alex Sim[2,]
Robert Drach[3]
Phil Jones[4]
David Brown[5], Julien Chastang[5], Luca Cinquini[5], Peter Fox[5], Danielle Harper[5], Nathan Hook[5], Eric Nienhouse[5], Gary Strand[5], Patrick West[5], Hannah Wilcox[5], Nathaniel Wilhelmi[5], Stephan Zednik[5]
Steve Hankin[6], Roland Schweitzer[6]
David Bernholdt[7], Meili Chen[7], Jens Schwidder[7], Sudharshan Vazhkudai[7]
S. Bharathi[8], Ann Chervenak[8], Robert Schuler[8], Mei-Hui Su[8]

Climate simulation data are now securely accessed, monitored, cataloged, transported, and distributed to the national and international climate community

[1] Argonne National Laboratory, Chicago
[2] Lawrence Berkeley National Laboratory
[3] Lawrence Livermore National Laboratory
[4] Los Alamos National Laboratory
[5] National Center for Atmospheric Research
[6] National Oceanic and Atmospheric Administration
[7] Oak Ridge National Laboratory
[8] University of Southern California, Information Sciences Institute

Table of Contents

1 Abstract

To meet the growing global demand for resources in computational climate science, the Earth System Grid SciDAC Center for Enabling Technologies (ESG-CET) aims to provide tens of petabytes of data, models, analysis, and visualizations, as well as the computational capabilities useful to access, manipulate, process, analyze, and interpret those data. These services are provided to a wide variety of stakeholders, including scientists, policymakers, and the general public. In addition, ESG-CET collaborates with other institutions and universities in researching data management, visualization, and workflow analysis. The goals of the ESG-CET project are to:

- Make data more useful to climate researchers by advancing Grid technology;
- Meet specific needs of national and international climate projects for distributed database, data access, and data movement;
- Provide secure web-based data portals giving access to broad-based multi-model data collections; and to
- Provide international climate centers and US government agencies with a wide-range of Grid-enabled climate data analysis tools and diagnostic methods.

2 Introduction

Sponsored by the Department of Energy (DOE) Scientific Discovery through Advanced Computing (SciDAC)-2 program, through the Offices of Advanced Scientific Computing Research (OASCR) and the Offices of Biological and Environmental Research (OBER), the Earth System Grid Center for Enabling Technology (ESG-CET)[1] is a collaboration of seven US research laboratories and one university. Together, these organizations are working to identify, create, and implement key computational and informational technologies that will advance the state-of-the-art in the science of climate change. We continue to support and expand operation of production "science gateways", on which the world-wide climate modeling community has come to depend, as well as designing, developing, and deploying the next generation ESG to support the needs of our stakeholders.

Since production began in 2004, the Earth System Grid (ESG) has housed and distributed significant and often extremely large data collections for many well-known efforts in climate science. As of April 2009, the ESG production system has over 14,000 registered users. ESG manages approximately 237 TB of model data, comprising the contents of archives at five sites around the United States. ESG users have downloaded more than 700 TB of data.

ESG users employ these data for a wide variety of purposes. The most notable example is that over 500 peer-reviewed scientific publications have been authored based on analyses of data in ESG's Coupled Model Intercomparison Project, Phase 3 (CMIP3) data archive. Some of these works contributed to the Nobel Prize-winning IPCC Fourth Assessment Report (AR4).

[1] http://esg-pcmdi.llnl.gov/

Thus, the Earth System Grid, first put into production under the previous SciDAC-1 project, and which continues to be operated by the ESG-CET project, has become widely known as a world-class source for climate modeling data through its support for important modeling activities in the US and around the world (see Section 3).

The success of the ESG in facilitating widespread collaboration through enhanced availability and accessibility of climate modeling data, combined with the growing recognition of the importance of modeling and simulation to a thorough understanding of and formulation of responses to our changing global climate, are the fundamental driving forces behind the Center's second area of activity, scaling the ESG to meet new and expanded needs.

More specifically, we are working closely with a number of prominent stakeholders to define and develop the next generation ESG system:

- The Coupled Model Intercomparison Project, Phase 5 (CMIP5) for scientists contributing to the IPCC Fifth Assessment Report (AR5),
- The Community Climate System Model (CCSM), primarily through the SciDAC-2 climate application project "A Scalable and Extensible Earth System Model for Climate Change Science", and
- The Computational Climate End Station (CCES) at the ORNL Leadership Computing Facility

In addition to these groups, which were anticipated and discussed in our proposal, several other groups have emerged as important partners in the future of the ESG:

- The Global Organization of Earth System Science Portals (GO-ESSP) is a forum for international cooperation around portal development and deployment, and
- The North American Regional Climate Change Assessment Program (NARCCAP) is a large collaborative project which has adopted the ESG as its data sharing and publication infrastructure

In order to address the needs of these stakeholders and the larger climate modeling community, we are developing a "next-generation" ESG environment to support a loose federation of more numerous and more widely distributed contributors to the data archives, with the majority of sites outside the immediate control and management of ESG-CET participants. An important characteristic of the next generation ESG is that it will provide users with only the information they truly require, and will avoid inundating the user in data. This goal will be accomplished by incorporating server-side analysis capabilities that reduce the raw data before it is delivered to the user. Creating this sophisticated capability requires broadening ESG to support multiple types of model and observational data; providing more powerful client-side ESG access and more sophisticated analysis services; enhancing interoperability between common climate analysis tools and ESG; and enabling end-to-end workflow of simulation and analysis.

Our partners have set ambitious scientific goals, which, in turn, pose some demanding challenges for data access, manipulation, analysis, and display. ESG-CET is actively innovating in many aspects of computer science and data management in order to deliver the tools the partners require to succeed. Before the end of the SciDAC2 project, for example, the ESG will likely be used to offer tens of petabytes of data across a federation of as many as two-dozen data centers worldwide. ESG will have to provide seamless access to these data, both to web-based clients and directly to analysis clients. Figure 1 depicts the projected evolution of ESG from 2006 through 2011. It indicates the scientific data

management and analysis requirements that have been and still must be met as the system matures. One specific point of interest is that the incorporation of all the features the figure describes through "Early 2009," will enable the establishment of a distributed test bed for CMIP5 (IPCC AR5) by mid-2009. The current status of this work is described in greater detail in Section 4. Appendix A and Appendix B describe the extensive collaborative activities carried out in conjunction with this work, and our community outreach through publications, presentations, and operational data portals.

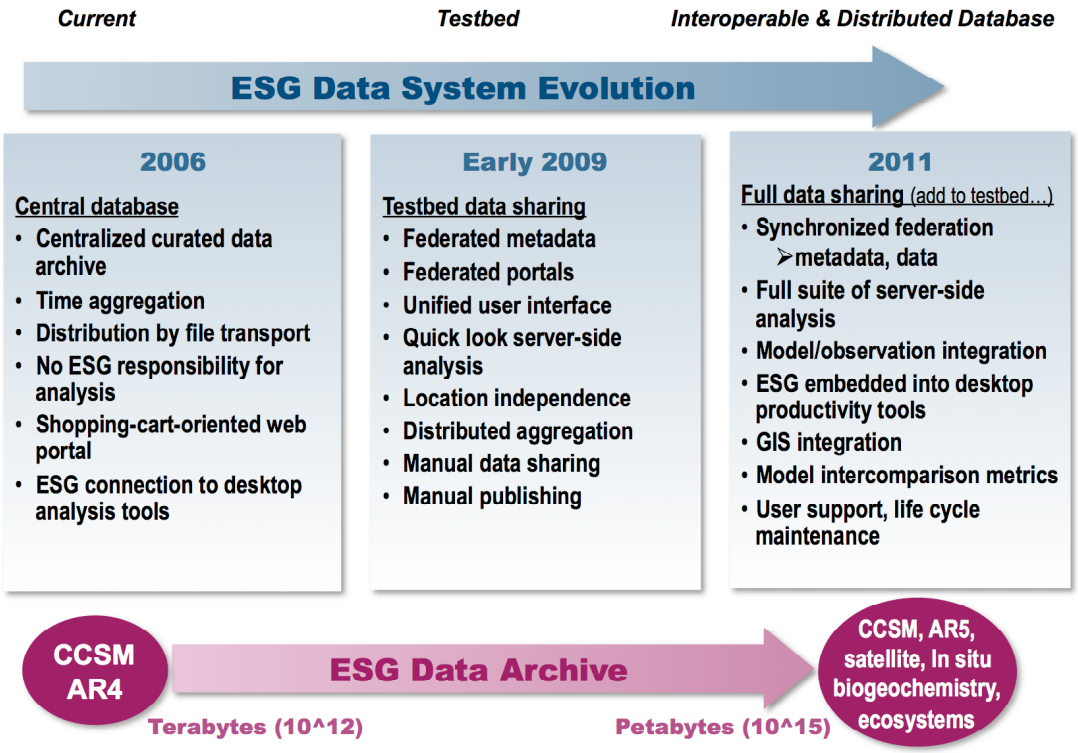


Figure 1. Evolving ESG to the Petascale: High-level ESG-CET Roadmap.

The project's annual and total budget, along with current staffing levels, is shown in Table 1.

Table 1. Project budget (annual and five-year total) and current staffing levels.

Institution	Yearly Funding	Total Funding Over Five Years
ANL	$460K [1.75 FTEs]	$2,300K
LANL	$170K [0.5 FTE]	$850K
LBNL	$320K [1 FTE]	$1,600K
LLNL	$480K [1.35 FTEs]	$2,400K
NCAR	$560K [3.4 FTEs]	$2,800K
ORNL	$350K [1.25 FTEs]	$1,750K
PMEL	$180K [1.1 FTEs]	$900K
USC/ISI	$230K [0.9 FTE]	$1,150K
Total	**$2,750K [12 FTEs]**	**$13,750K**

3 Overall Science Impact Highlights

In our ESG-CET SciDAC2 proposal, sustaining and expanding our operational capabilities based on the existing (and lightly augmented) infrastructure was called out as one of our top priorities. In parallel with the development of our next-generation system, ESG-CET has continually invested substantial resources in providing operational services for an international climate change science community. In the sections that follow, we summarize the overall progress and impact associated with this component of our project.

3.1 Intergovernmental Panel on Climate Change (IPCC) 4th Assessment Report (AR4)

The World Climate Research Programme (WCRP) assists scientists in addressing key scientific questions regarding climate change through its Coupled Model Intercomparison Project, Phase 3 (CMIP3). Through its CMIP3 Multi-Model ESG data portal, hosted at PCMDI, the WCRP has cataloged and archived 23 global coupled ocean–atmosphere models for the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC). (The collective scientific achievement of the IPCC was recognized in 2007 with the award of the Nobel Peace Prize.) The most frequently downloaded climate data archive since 2005, the CMIP3 Multi-Model archive serves more than 3,000 registered groups that together have downloaded over 600 TB of data. The archive contains over 35 TB of data, and its data have been downloaded at a rate averaging 500 GB per day (***Figure 2***). To date,

more than 500 publications[2] have appeared that expound scientific results based on data from the CMIP3 archive.

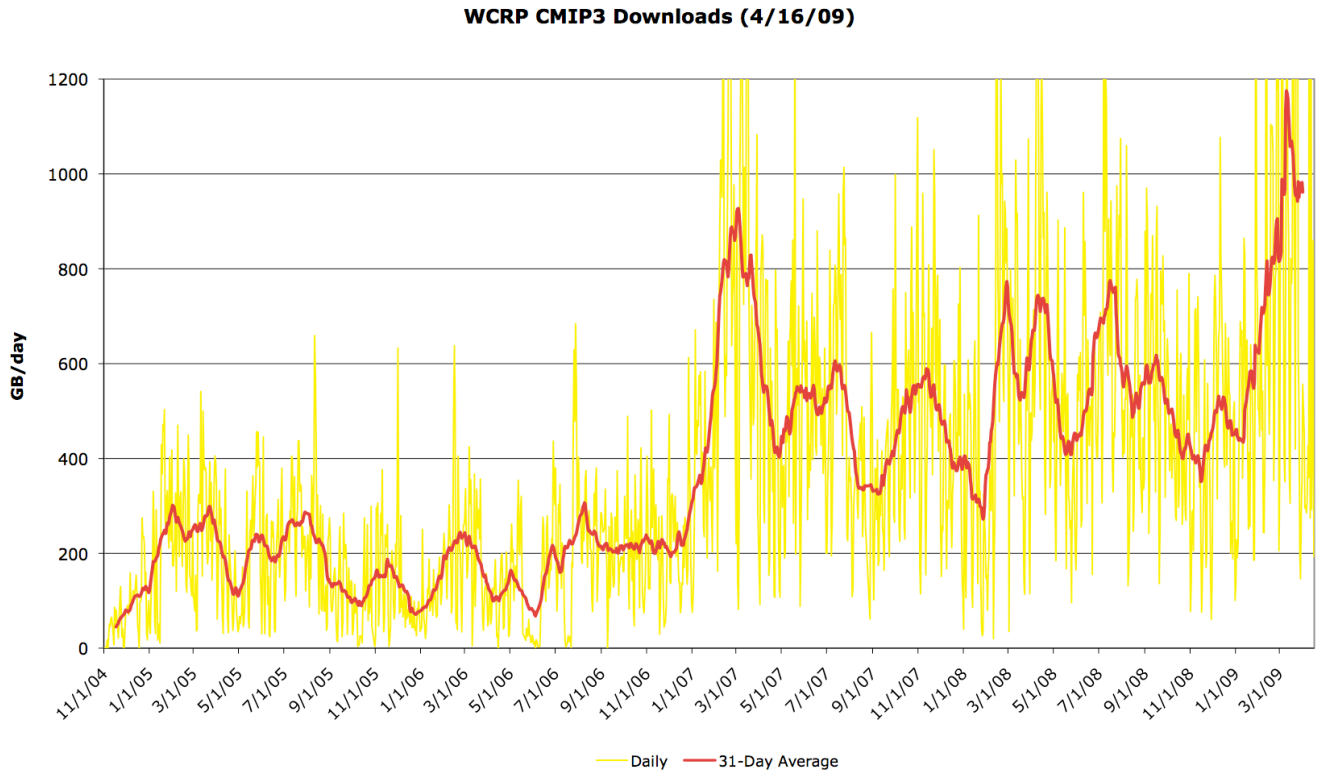


Figure 2. Amount of data downloaded from the ESG archive. Daily download quantities are plotted in yellow, with a 31-day running average superimposed in red. On some days the amount of data downloaded has reached 1 TB.

3.2 Community Climate System Model (CCSM)

The ESG-CET CCSM and climate research portal, www.earthsystemgrid.org, serves as a "science gateway" for US based climate research, and is operated at NCAR. The portal provides a range of offerings that spans CCSM results alongside the Community Land Model (CLM), the Parallel Climate Model (PCM), the Parallel Ocean Program (POP), the Community Land-surface Model (CLM), the Community Sea Ice Model (CSIM), the CCSM model source code and related files, as well as analysis and visualization tools. As of March 2009, we were providing almost 200TB of products to a community of over 13,000 registered users, and this user community has downloaded 93 TB of these products using the tools and access provided by ESG-CET. These data are separate from the CCSM contribution to the CMIP3 archive, as they comprise many more experiments than those required by the IPCC for its Fourth Assessment Report. Of particular interest to the ESG-CET community are the CCSM datasets that were saved at 6-hour time intervals; these data have proved extremely useful to the regional modeling community as well as to the climate impacts community (Figure 3). In addition, CCSM source code and input datasets have been downloaded over 2,400 times on their own.

[2] http://www-pcmdi.llnl.gov/ipcc/subproject_publications.php

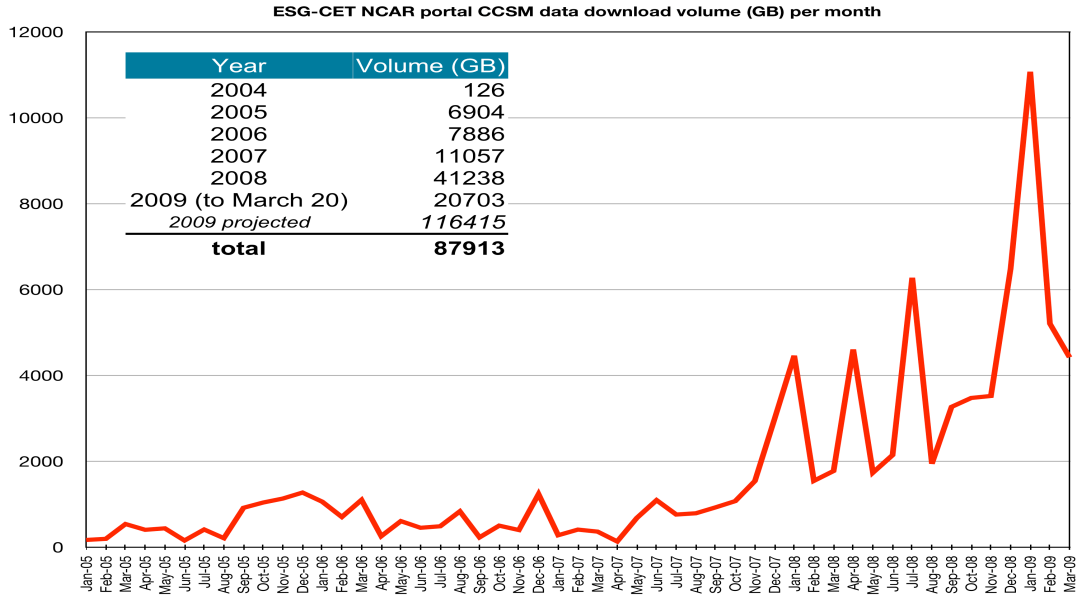


Figure 3. The monthly CCSM data download volume from the ESG-CET portal at NCAR continues to grow over time, as more members of the community find value in the ESG-CET CCSM data holdings. The anomalously high download total for January '09 is due to 6-hourly CCSM (IPCC AR4) data becoming available to regional climate modelers.

3.3 Carbon Land Model Intercomparison Project (C-LAMP)

The ESG-CET team has deployed a standalone portal at Oak Ridge National Laboratory (ORNL) in support of the Carbon Land Model Intercomparison Project (C-LAMP), a project designed to enable the international scientific community to evaluate the performance of biogeochemical models normally coupled to general circulation models. Supported in part by the SciDAC CCSM project, the C-LAMP effort uses the resources of the Computational Climate End Station (CCES) at the ORNL Leadership Computing Facility. Both the SciDAC CCSM and CCES at ORNL are key stakeholders in the ESG. Through the C-LAMP ESG portal, model data from C-LAMP experiments are made available for detailed analysis by the developers of biogeochemical models. The archive, holding 149 GB of data in 1575 files, currently has nine registered users that represent C-LAMP's core model-development team and members. C-LAMP plans to publish significantly more data in the coming months, including several terabytes of diurnal statistics and additional atmospheric model outputs; the project also plans to expand access to accommodate additional users.

4 R&D for the Next-Generation ESG-CET

As anticipated in the proposal, a significant redesign of the ESG was necessary in order to address shortcomings of the previous infrastructure, as well as to meet emerging scientific challenges and embrace new community and industry developments. Some of the most important considerations that factored in the redesign are:

- the massive amount of output expected from the next generation of global and regional models makes it impossible to transfer and serve all data from a single location, and suggests instead to serve the data from the location it was generated;

- the emergence of thematic projects and communities (e.g. CMIP5/IPCC-AR5, CCSM, CCES, NARCCAP) requires a flexible approach to the development of data services, where a common layer of services and infrastructure must be combined with custom user interfaces (UIs), workflows, and domain specific end-user applications;
- each data center is mandated to support its mission critical operations based only on its own resources, without having to rely on external parties;
- the increasing number of stakeholders (both nationally and internationally) in the geo-scientific data distribution environment mandates a collaborative approach, based on commonly agreed interfaces for interoperating among partners data centers, as well as the development or adoption of technologies for federating among virtual organizations; and
- finally, the continuous development of sophisticated server and client side applications by the community (e.g. TDS, LAS, CDAT, NCL, Ferret) as well as the increasing adoption of common metadata standards and protocols (e.g. CF, OPeNDAP, OAI) presents an opportunity for integrating mature and well adopted technologies, as opposed to undertaking the impossible task of developing everything from scratch.

The goals for the next-generation ESG design from the user perspective were to provide a seamless experience, in which rich search capabilities would allow the identification of data of interest throughout the widely distributed ESG holdings, and allow requests to be made for anything from the raw data to a variety of analysis and visualization products. From the perspective of data providers and projects, the goals are to be able to flexibly publish and manage access to data collections, as well as the possibility of offering post-processing services tailored to the project. From the perspective of the data centers hosting ESG services, there is a need to simplify participation and minimize the administrative burden, while insulating the overall ESG enterprise from too strong an interdependence on resources contributed by others.

4.1 Overall Architectural Design

In order to satisfy the various needs outlined above, the architecture of the next generation ESG system is designed to be much more geographically distributed, federated across independent data centers, and based on the integration of new and existing components (see Figure 4 and Figure 5). Logically, the system is composed of 4 separate layers.

Discipline specific *Gateways* represent the main entry points for users (and user clients) into the system. A Gateway is a web-accessible application that includes all UI elements exposed to the user, as well as high-level data services (for data browsing, search and discovery), user administration and access control management. Gateways may also act as a broker for data requests sent to the Data Node application servers. Logically, a Gateway enables a user to access data from one or more associated Data Nodes. But since Gateways exchange metadata with each other, and are federated through a common security infrastructure, a user can seamlessly find and access data that is served from anywhere in the system.

A system of *Data Nodes* serves data and data products securely and reliably to the end users. Data Nodes are application servers co-located with the data, which can be flexibly configured depending on the specific community needs and hardware and administrative resources. Data Nodes can offer one or more services such as direct file access via an HTTP/FTP/GridFTP server, deep storage data retrieval via BeStMan, data sub-setting capabilities via OPeNDAP, and higher-level post-processing and visualization capabilities via a LAS Data Product server in combination with an analysis engine such as CDAT, Ferret, or NCL.

A thin layer of *Global Services* supports the interoperability among virtual organizations without impeding their mission critical operations. This layer (which is still in the design phase) will most likely include a global registry for Gateway endpoints, access control attributes, and metadata namespaces, as well as some general infrastructure for monitoring the various components of the system and notifying the system administrators.

A *Client Applications* layer provides end users (data users and data providers) an interface to access the heterogenous services offered by the system. Supported clients will include web browsers, desktop applications for data publishing and data movement, as well as 4th generation analysis and visualization languages such as CDAT, Ferret, and NCL.

In the following sections, we describe some of the key elements and cross-cuttings aspects of the next-generation ESG system, which is now being deployed in our initial testbed environment.

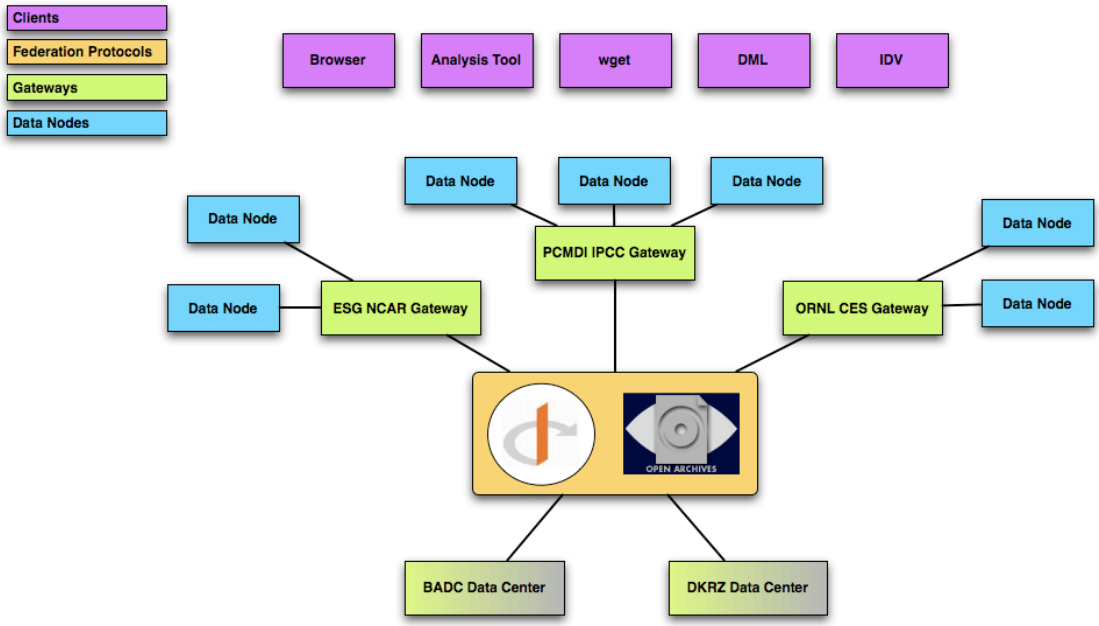


Figure 4. ESG-CET system topology, as envisioned by the end of year 2009. A network of geographically distributed Gateways and Data Nodes is welded into a global federation by the adoption of standard protocols and APIs (such as OpenID and OAI), thus allowing seamless access by a vast array of end-user client applications.

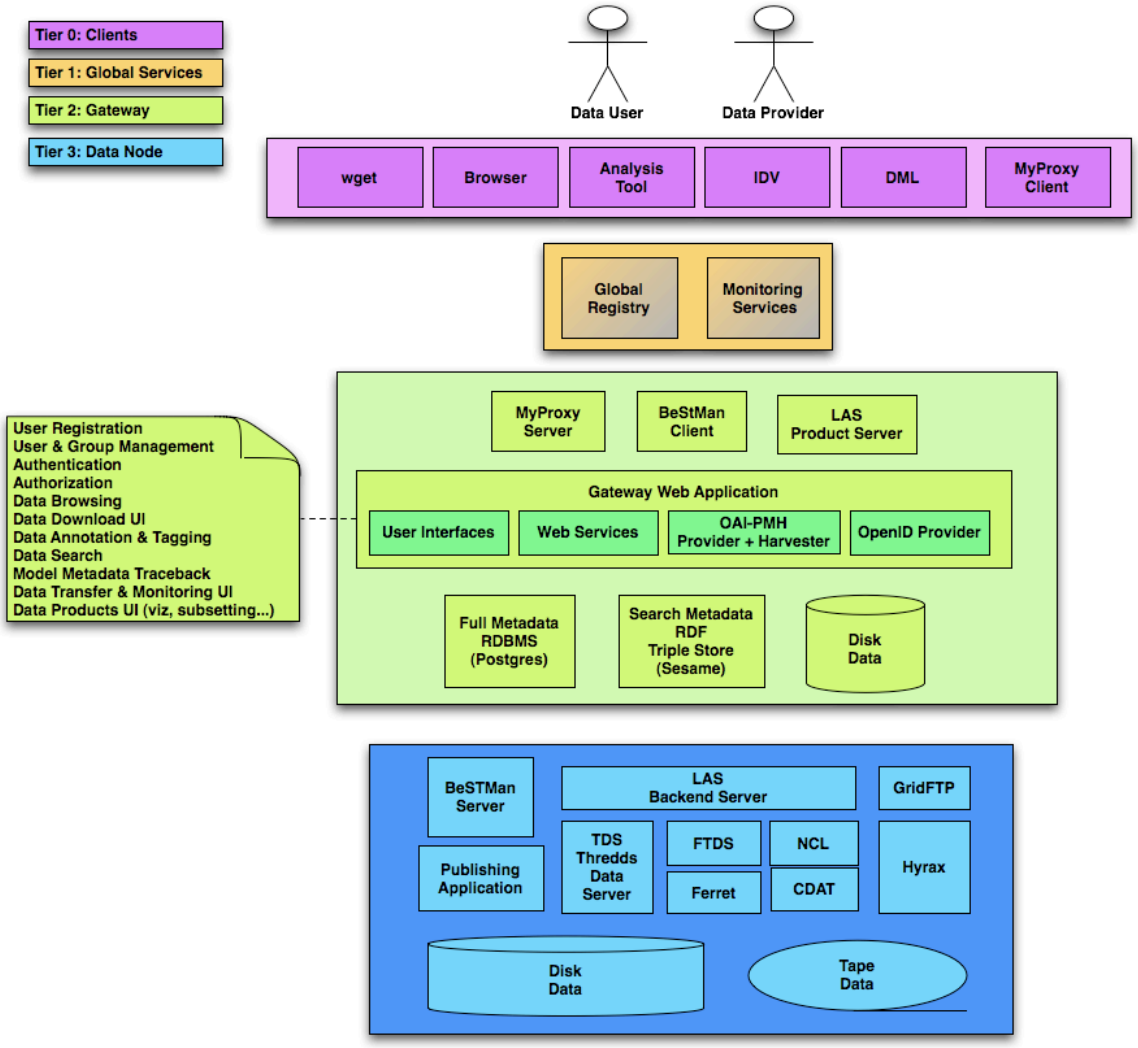


Figure 5. The ESG-CET software stack, structured into four logical layers, integrates new software components with well-established protocols, servers and clients that have widespread community acceptance.

4.2 Gateways

Gateways are the main entry points for users to access the data and services offered by the ESG federation. Gateways allow users to browse and search for data, examine detailed metadata, download and subset files, request high level data products such as analysis and visualization, and register and apply for specific group memberships. Local administrators also manage community groups at the Gateway. Additionally, Gateways are federated, so that services are available to the users seamlessly throughout the system, no matter where they first registered or where they started their working session. A Gateway is typically configured, customized and populated with resources to serve the specific needs of a project, an institution, or a community. Each Gateway acts as a broker to access the data holdings served by one or more associated Data Nodes, because it provides infrastructure to index the data, to authorize user access, to expose the data access points, and (in the case of higher level data products) to formulate and submit the data request to the Data Node on behalf of the user.

Internally, a Gateway is structured as a typical multi-layered web application (see Figure 4). Metadata persists in a combination of a relational database and RDF triple store. A stack of data access, service and control layers exposes the Gateway functionality to users and clients either as a suite of user interfaces, or as web service endpoints (SOAP or REST). In particular, a Gateway acts as an OAI-PMH Provider and Harvester to enable exchange of metadata records with other Gateways and partner Data Centers. Also, a Gateway may act as its own Identity Provider for web-browser user authentication, or it may be connected to an external Identity Provider. A MyProxy server allows users to obtain short-lived digital certificates that may be used by client applications to directly access data holdings from the Data Nodes in the system.

The current ESG-CET federation test bed is composed of two Gateways, at NCAR and PCMDI, as well as a number of prototype Gateways that were setup to serve specific communities and projects (e.g. CADIS, DyCore, Curator). It is expected that by summer of 2009 the currently deployed Gateways will reach full operational status, after which a number of other discipline specific Gateways (e.g. CES, CDP, etc.) will be deployed and will join the federation.

4.2.1 *User Interface*

Each Gateway within the ESG-CET federation provides a web-accessible user interface (UI) enabling user interaction with the services provided by the system. The Gateway user interface allows end users to search and discover scientific data from the entire federated system, browse data collection hierarchies from a discipline specific Gateway, download data collection files individually or in bulk, sub-set collection files, visualize data collections by variable, track deep storage file download requests and access user profile information.

User interface components have been developed for the key functional areas listed below. Significant progress has been achieved in each component area and the listed components will be included in the first release of the Gateway application. We expect to refine the appearance and behavior of each component as the broader user community uses the system. The design goals of the Gateway user interface include exposing the base system functionality, providing a consistent and intuitive user experience and supporting a flexible and maintainable framework for future enhancements and revisions.

- *Home Page*. The home page provides visitors with general information about the discipline-specific Gateway, starting points for discovering data collections, direct access to notable data collections, important notices regarding system status and access to login and account request functions. The home page is highly customizable, allowing a Gateway to present discipline specific information, data browse entry points, logo images, and color palette.
- *User Registration.* The user registration component offers a multi-step workflow for account creation, approval and validation. The resulting account may be used to authenticate any user at any Gateway in the federation.
- *User and Group Management.* The user and group management component provides tools for registered users to change account settings, request access to privileged data collections and tools for group administrators to approve group requests and manage group membership.
- *Login.* The login component allows registered users to authenticate with the federated Gateway system with an OpenID user identifier. Users may request their password delivered in email in case of lost credentials or make inquiries to Gateway support staff.

- *Data Browsing.* Data collections may be viewed in a variety of ways in the data-browsing component. Support for file system-like hierarchies and high-level associative arrangements such as experiment and project related listings are provided.
- *Data Search.* Data collections may be searched in a variety of ways. The search component provides simple and familiar text-based search as well as advanced capabilities using a dynamic faceted query interface for detailed metadata inquiry.
- *Data Download.* Data collection files may be downloaded individually via hyperlink or requested in bulk from the file-listing interface using generated "*wget*" scripts. Often these data collections are restricted and under access control in which case the user is directed to authenticate prior to data download.
- *Data Transfer.* The data transfer component allows registered and authorized users to request and manage groups of files from deep storage systems throughout the federation. Users can access real time status reports and are notified by email when transfers are complete.
- *Data Annotation.* Registered users may add comments and keywords to data collections and files. This enables the user community to further describe and enhance Gateway content. The group administrator may impose a review process for particular collections.
- *Data Visualization and Sub-Setting.* The data visualization component provides an interface for requesting charts, plots and data sub-set downloads. Users may choose variables of interest, select sub-regions geo-spatially with an interactive map and temporally with time controls.
- *Model Metadata Trackback.* The Model Metadata Trackback component allows detailed inspection of the metadata associated with Models, Model Components and Model Component Frameworks (including model sub-components, grid specifications, scientific configuration, and model run input.) This interface was developed in collaboration with the Earth System Curator project.

  Currently, the only information on the component framework is the name, if there is one present. The focus has been on describing Models, Model Components, and Simulations (configured, executed models) in as generic a way as possible.
- *User Workspace.* The workspace component is meant to address the specific scientific needs of individuals and groups. It is been developed as a virtual space where users and groups of users can save resources of interests (datasets, models, simulations, etc.) for analysis, comparison, and sharing with colleagues.
- *Metrics Reporting.* The metrics reporting component allows registered and sufficiently privileged users to generate formal and ad-hoc reports detailing Gateway use information. These reports can be generated graphically or texturally and may be delivered automatically by email.

4.2.2 ***Registration***

The ESG System features a flexible and federated model for registering users and assigning them special group memberships and roles, that are used both to establish access control restrictions on collections of resources, and to gather any user information that may be requested by either data providers or funding agencies (see Figure 6).

In general, a user may visit an ESG Gateway as a guest, i.e. without authenticating. In this case, the Gateway has no information about the real identity of the user, and allows only limited functionality to

be invoked. For example, data browsing and searching are free operations, while downloading of actual files is typically restricted to authenticated users.

In order to obtain access to a wider range of data and services, a user may elect to establish an ESG account by completing a registration process. ESG registration involves only a few mandatory fields (first name, last name and email), in addition to a few optional fields, and requires the user to validate their email before the account is activated. After the process is complete, the user may log into the ESG Gateway where he/she registered, and will typically be able to access more services and products. For example, many datasets are restricted to authenticate users because the data owners want to be able to know the identity of whoever has downloaded the data (for notification or reporting purposes, for example).

Additionally, at the time of registration the users are assigned an OpenID – a URL that uniquely identifies the user across all Gateways (i.e., within the ESG federation). When the user visits another Gateway, he/she can use that same OpenID to login at that Gateway, through a redirection process that will eventually have the users enter their credentials at the Gateway where they originally registered. Thus by completing one registration process, the user is able to authenticate with any Gateway in the federation.

Some resources hosted on an ESG Gateway may require additional privileges to be accessed by a user, besides those granted by authentication alone. This may happen either because the data provider owning the resources is required to gather additional information (a statement of work, for example), or because the resources are not ready or intended for general access. In this case, access to the resources is restricted to members of a specific User Group, which is controlled and administered by one or more group administrators. A user may apply for membership in the group by providing the requested information, at which point the administrator is responsible for accepting or rejecting the user’s application, and for assigning the user one or more roles within the group.

Note that each group is associated with (a.k.a. "owned by") a single Gateway, which is the Gateway where all users across the federation will be redirected to when applying for membership in that group. This model allows a group to be managed by its administrators on a consistent user interface, and to keep track of all the group's users and user attributes in a single location.

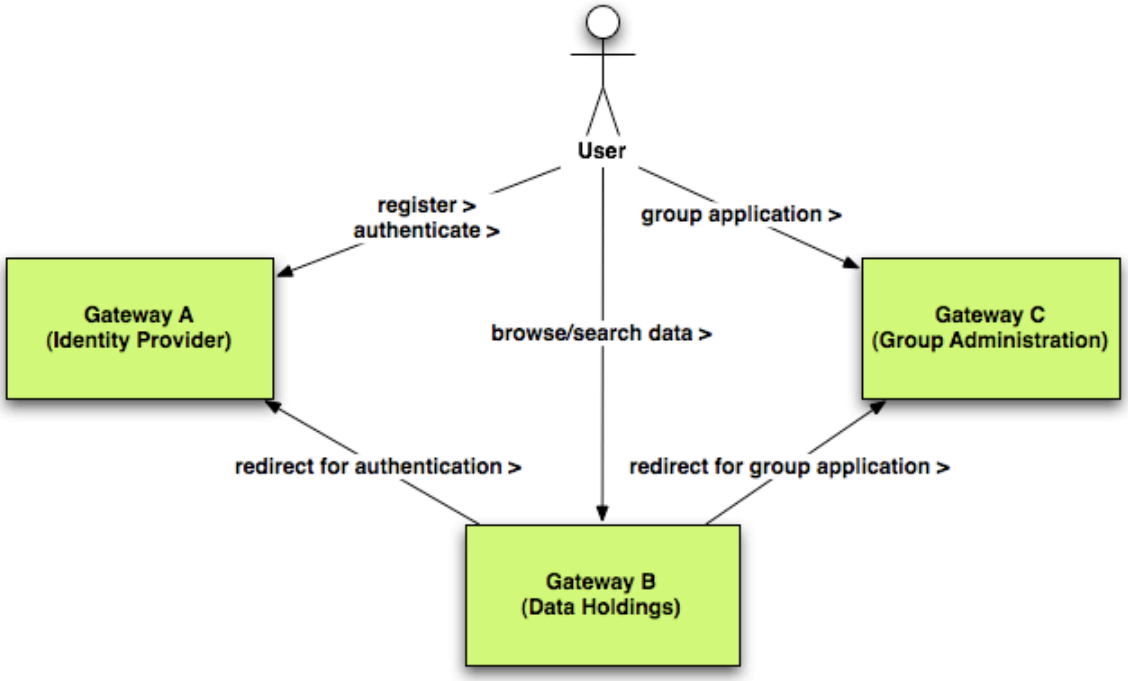


Figure 6. The ESG federated architecture separates the registration/authentication, data access and group affiliation functionality into three separate logical components that may be hosted on distinct Gateways.

4.3 Data Nodes

The Data Node, together with the Gateway, forms the two main components of the ESG architecture (see Figure 4). Data Nodes are hosts where data is actually stored or archived. As Figure 7 illustrates, the functionality of the Data Node is to publish data (making it visible to a Gateway), and provide data analysis and delivery services to ESG end users.

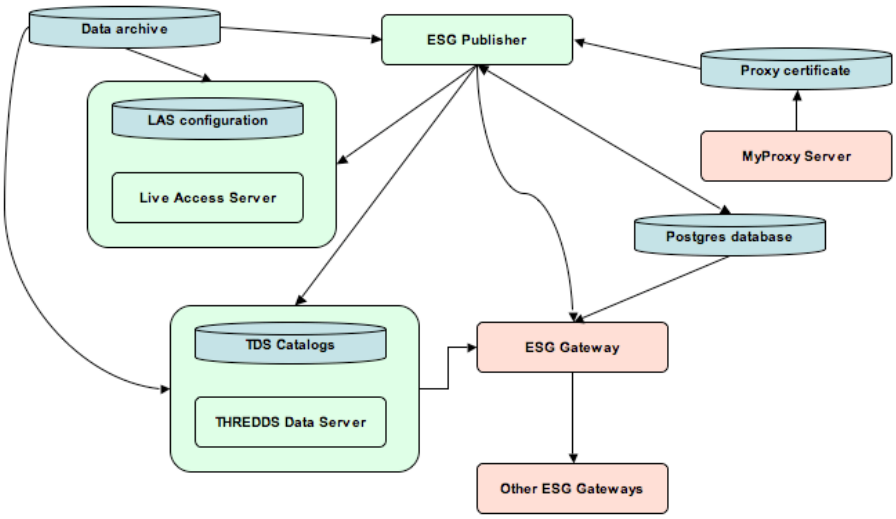


Figure 7. The above diagram indicates Data Node components in green, data and configuration files in blue. The red components are the subset of Gateway services that interact with the publisher.

Data Node services include:

- ESG Publisher, for scanning data, generating catalogs and data aggregations, and publishing to the Gateway portal;
- THREDDS Data Server, to serve data and data aggregations with the http protocol;
- Live Access Server, for creating visualization and data subset products;
- Data Mover Lite, for bulk data transfer; and
- GridFTP, to serve data through the GridFTP protocol.

Data Node components can flexibly handle data from rotating disk, tertiary mass storage archival systems, and even data stored in databases. Most data is expected to be in gridded form; datasets conforming to the netCDF CF-1 metadata standard can be interpreted by all Data Node components. At present, we have built testbed ESG Data Nodes at PCMDI and NCAR. The PCMDI Data Node has successfully published the 35 TB CMIP3 climate data archive.

Integrated into the backend of LAS "Product Services" is the Climate Data Analysis Tools (CDAT) that performs comparison, analysis, and visualization of large model output datasets, and delivers results to climate scientists as customized end products accessed through a web browser. With CDAT analysis capabilities in LAS, climate scientists on the ESG grid can transparently access large datasets from disparate data centers and perform complex, configurable, ESG "Data Node" operations in a lightweight user interface environment.

Ferret comes standard with LAS. Thus, Ferret is already an integral part of the LAS "Product Services." Next to be integrated into the backend of LAS is NCL. Once all three-analysis engines are integrated and test with the "Product Services", a more standard API will be developed to easily incorporate other commonly used analysis tools such as GrADS, IDV, etc.

4.3.1 *Data Publishing*

Data publication is the process of making data visible for search and download from an ESG Gateway. The application that does this is the "ESG Data publisher". The publisher runs on a Data Node (where the data resides) and has several functions: it scans and extracts metadata from the source data files; it stores that information locally and sets up catalogs for other node processes; it communicates that information to the ESG Gateway.

The publisher coordinates the main functions on the Data Node. It integrates with the THREDDS Data Server that actually serves data files and aggregations of data. It interacts with the Live Access Server (LAS) that produces graphics product, ensuring that newly published data is made known to the LAS. Most importantly, it signals the Gateway that new data products are available for download and analysis. The client supports single sign on to portal functions with the MyProxy service. MyProxy generates limited lifetime certificates, modified with ESG-specific attribute information. The publisher serves as the coordinator of these processes, and stores relevant metadata locally, so that data providers can query the published information easily.

Testing the publisher component is currently underway on the two ESG testbed nodes at PCMDI and NCAR. The publisher has been ported to several programming environments: Linux, Sun Solaris, and Mac OS X. The node environment will be replicated at many ESG production sites with widely varying environments. The publisher is highly configurable, and can run in graphical user interface (GUI) mode or from the command line. In either mode the data provider can publish one or multiple datasets at a

time. We have tested the publisher by processing the entire CMIP3 archive used for the IPCC AR4 report, an archive consisting of over 1000 climate model datasets containing 35 TB of data. This is representative of the scale of data to be published in ESG.

4.3.2 *OPeNDAP-G*

At the heart of the Earth System Grid is the need for access and transport of various data products. These products are produced and consumed in many and varied parts of the ESG architecture, including on Data Nodes, Gateways, clients and federated sites. The NCAR/HAO effort involves secure, high-performance access and transport of data within the ESG-CET framework.

To date, we have migrated all functionality (except aggregation) of the ESG-specific add-ons to OPeNDAP-G into the core of OPeNDAP/Hyrax and this software has been available for public use since 2006. This software supports multi-dataset access into a single OPeNDAP object, using the official version of the netCDF module instead of the ESG-specific module. All OPeNDAP-G code for ESG is up to date with the current 1.4 Hyrax release software. (Hyrax 1.5 is pending release.) We are working to use the new fileout_netCDF module from OPeNDAP, which was taken from the ESG version. This work involves new service capabilities, and XML document support in requests and responses.

We have developed a client library that communicates with an OPeNDAP Hyrax Back-End Server (BES) module that allows the client to make netCDF library calls on remote data. We also developed a GridFTP module that allows these netCDF calls to be passed through a GridFTP server that performs user authentication using the ESG security infrastructure. The GridFTP protocol support is implemented for the Globus 4.x releases. The RNI (Remote netCDF Invocations) provides multi-tier gridded netCDF calls.

To support flexible configuration of the architecture and ESG's need for adaptable front-end services, we created pptcapi (OPeNDAP PPT C API) to be used by C-language components of ESG, specifically RNI.

For the ESG-CET security infrastructure, we provide credential support between ppt/pptcapi to the OPeNDAP BES, specifically X509 PKI credentials passed between client and server. Further, all OPeNDAP software is security audited using the Fortify code analysis tool as well as subjected to all OPeNDAP security policies and procedures. All developed software has full build, configure and test suites and are deployable on all ESG-CET hardware/software systems.

4.3.3 *Product Services and the Live Access Server (LAS)*

Among the institutions making up ESG-CET is the Pacific Marine Environmental Laboratory (PMEL) of the National Oceanic and Atmospheric Administration (NOAA). At the heart of PMEL's contribution to the ESG-CET lies the Live Access Server (LAS), an XML-configurable workflow engine (under development since 1993) that has supported widely varied climate projects. The incorporation of LAS into the ESG in 2006 brought the accomplishments of those projects to the ESG project; it also made ESG-specific enhancements to LAS available to those projects.

The ESG-CET is intended to serve information products to users representing a broad spectrum of sophistication -- from numerical modelers who want access to "raw" model output files and verbatim subsets of model output; to climate impacts investigators who want rapid access to these data without the complexities of model-specific coordinate systems; to those users who only want to quickly visualize the overall behaviors of models. The petascale nature of the ESG data holdings require that significant levels of data reduction take place at the server in order to satisfy these customers – both through

straightforward subsetting and decimation and through specific analysis operations, such as the computing of spatio-temporal averages. In the ESG-CET architecture, we refer to the steps that convert raw data into analysis results and visualizations as "product services." Architecturally, the product services lie above the OPeNDAP layer of ESG, which provides a service-level access to virtual datasets (aggregated files) through the netCDF-CF API. Here in bullet form are the highlights of the enhancements to LAS made through ESG support.

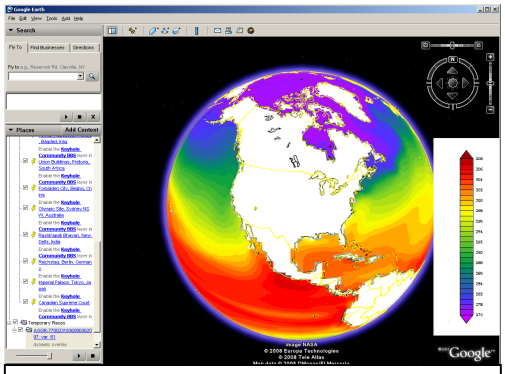


Figure 10. LAS output to Google Earth®

- Visualizations on Google Earth® that dynamically expand the contour levels (and optionally the color scale) as the user zooms in (Figure 10)
- Animations created on-the-fly that are customizable in time and space range as well as graphical style
- An API and library to support Ajax-based user interface development for LAS
- Server-side data reduction (averaging, regridding, binning, etc.) embedded into OPeNDAP requests, including leadership in community-wide discussions towards a standardized syntax for such transformations
- Joining LAS with the Climate Data Analysis Tool (CDAT) at DOE/LLNL/PCMDI, so that CDAT products can be obtained through LAS. (Similar work is underway collaborating with NCAR to blend LAS with NCAR's NCL package.)
- The ability to incorporate observed data (points, profiles, time series, etc.) into LAS data products through the DAPPER and ERDAP protocols (profiles of OPeNDAP suitable for observation data)
- A tool called the "SlideSorter" that provides the ability to compare model fields and ensemble members including visualizations of difference fields (Figure 11)
- Support for curvilinear model grids through a visualization style that reveals the native grid structure, and automated configuration tools for such datasets
- A Web-based user interface allowing administrators to configure and manage LAS and its product cache

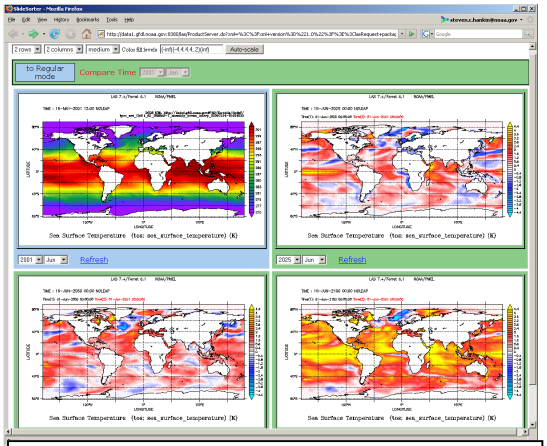


Figure 11. LAS Slide Sorter output

In particular, ESG enhanced the model-intercomparison capabilities of LAS, which has benefited the dissemination of operational global ocean circulation model outputs from the Hybrid Coordinate Ocean Model (HYCOM), a project DOE has supported with past funding. HYCOM LAS data services are

available on the internet.[3]. Similarly, ESG enhancements have benefited the seasonal-to-interannual climate forecast community, through the Global Ocean Data Assimilation Experiment (GODAE), with a number of LAS servers installed internationally[4] including the installation at the US GODAE Server.[5]. ESG developments have also had positive impacts on the ocean-climate observations community through LAS. The enhancements made to LAS in support of the ESG data publications process have found their way into the Observing System Monitoring Center (OSMC)[6] and related ecosystem monitoring activities.[7] As ESG-funded enhancements are added to LAS there will continue to be positive impacts across a range of significant projects that utilize the software.

4.3.4 *Data Reduction and Analysis*

As data sets become larger, ESG users will be performing more analysis remotely and will be relying on ESG to perform more than simple sub-setting and aggregation of data for downloading. Much of this analysis can be performed using tools like CDAT or Ferret that can provide a familiar front-end supported by the Live Access Server (see above). However, large data sets will require parallel analysis tools to provide rapid turnaround and interactive analysis of data. For example, in eddy resolving simulations of the ocean performed by LANL researchers at ORNL's Leadership Computing Facility, each instance of a 3-D field is 1.6 GB in size, so manipulating such fields or combinations of fields for analysis requires substantial computation and memory. We have installed the parallel visualization tool ParaView at both LANL and ORNL and have demonstrated its effectiveness at performing remote analysis of eddy-resolving ocean simulation data at LANL using resources at ORNL where the data resides. We are now in the process of integrating ParaView into ESG. Requirements for interaction with the Live Access Server have been developed and the design for integration is under way. We are also in the process of installing a new multi-processor Data Node at LANL data for LANL-based eddy-resolving ocean data.

In order to perform comparisons between models and data that reside on different meshes, interpolation is required. Many in the climate community use the Spherical Coordinate Remapping and Interpolation Package (SCRIP), developed and distributed by LANL. The current version of SCRIP uses a custom description of each grid and has some problems with robustness, particularly for grid points near the polar singularities. It also computes regridding interpolation weights off-line in a serial computation, leading to memory issues and long computation times for high-resolution grids. We are developing a next generation SCRIP to address these issues. First, we have developed a new algorithm for SCRIP that will greatly improve its robustness and are nearing completion of its implementation in the SCRIP software. In the process, we have also refactored SCRIP to provide a subroutine/component interface and enable the definition of grids distributed across processors or nodes. Finally, the grid description in the new SCRIP is based on an emerging climate grid specification standard. Adoption of this grid standard within ESG will greatly improve intercomparisons between climate model data.

---

[3] http://hycom.coaps.fsu.edu/las/getUI.do

[4] http://www.godae.org/Ocean-products.html

[5] http://www.usgodae.org/las/servlets/dataset

[6] http://osmc.noaa.gov/Monitor/OSMC/OSMC.html

[7] http://www.pfeg.noaa.gov/products/las_5_1.html

Once the new ESG architecture and basic subsetting and aggregation tools have been deployed, we will begin work on more advanced analysis tools. In particular, some "standard" fields in ocean data analysis require the computation of streamfunctions or transports that are not always aligned with the computational grid, but must be computed on those native grids to preserve flow properties. We will be developing flexible tools for performing such analyses.

4.4 Metadata

Not surprisingly, accurate and extensive metadata is absolutely critical to the operation of the ESG system as a whole, and to the interplay of the several ESG data services (see Figure 8).

A Domain Object Model – an abstract representation of all the entities and relationships in the system, represents the core of the ESG metadata. This model can be logically divided into four connected sub-domains:

- Scientific Metadata – metadata that is needed to describe and analyze scientific data. This includes concepts such as variables, grids, models, aggregations, and the associated scientific activities that produced the data such as projects, experiments, simulations, campaigns etc. Part of this metadata is harvested automatically from the data itself at the time the data is published, and part is inserted manually by the data providers based on their knowledge of the process that generated the data;
- Location Metadata – metadata that describes the physical location of the data (including replicas), and the services appropriate for accessing the data;
- Access Control Metadata – metadata that is used to authorize a user to execute an operation on a resource (for example, download a file or request a visualization product). This metadata includes user credentials, access control statements on resources or collection of resources, and information on user affiliation with specific communities that are granted access to the resources; and
- Metrics Metadata – metadata that captures the usage of the system by the user community, for reporting, usability and auditing purposes.

Metadata is captured and enters the system via the publication process, and persists in a Relational Database that maps the abstract Domain Object Model into a set of tables and table constraints. The relational database is the authoritative source of the ESG metadata, and drives the Gateway services for data browsing, data description, and access control.

Additionally, the decision was made to represent a limited subset of the full Domain Object Model as RDF/OWL Ontology, to drive the ESG search and discovery services. This choice was motivated by the flexibility of the semantically based approach, which allows easy configuration of different applications based on a common infrastructure (for example, the ESG and CADIS query services), and a straightforward extension to domain-specific functionality (like for example the metadata necessary to drive the Dynamical Core workshop). Because relational technologies still offer superior scalability, the ESG metadata architecture relies on harvesting only a limited portion of the full relational metadata (those fields that are important for the high level description of the data collections) into RDF triples and the resulting RDF Triple Store is then queried by the search and discovery services.

Finally, the ESG system leverages the well-established OAI-PMH technology (Open Archive Initiative – Protocol for Metadata Exchange) to exchange search and discovery metadata between ESG Gateways, to interoperate with partner Data Centers (by importing their metadata records and exporting the ESG records), and to disseminate geo-scientific metadata to any interested Digital Library. Each ESG Gateway includes functionality for operating as an OAI-PMH Provider (whereby the local metadata records stored in the RDF triple store are serialized as XML and sent in response to HTTP REST-like

requests), and as an OAI-PMH Harvester (whereby XML records from other OAI-PMH Providers are converted into RDF and stored in the Triple Store). The resulting architecture can be described as a peer-to-peer framework for semantic metadata exchange, which enables a user to start from any ESG Gateway in the system, and locate records at other Gateways, or partner Data Centers.

The modeling part of the ontology in ESG is increasingly based on that being developed by the EU Metafor project. The Curator project is serving as a liaison between Metafor and ESG. Curator members at NCAR are implementing portions of the Metafor ontology in the ESG Gateway and working with ESG and Metafor to reconcile differences. Curator has focused initially on implementation of a detailed, hierarchical grid specification in the gateway. Curator has also added detailed metadata describing atmospheric dynamical cores, developed for a workshop at NCAR in which multiple dynamical cores were compared. These ontology additions are being integrated into search and browse features of the Gateway.

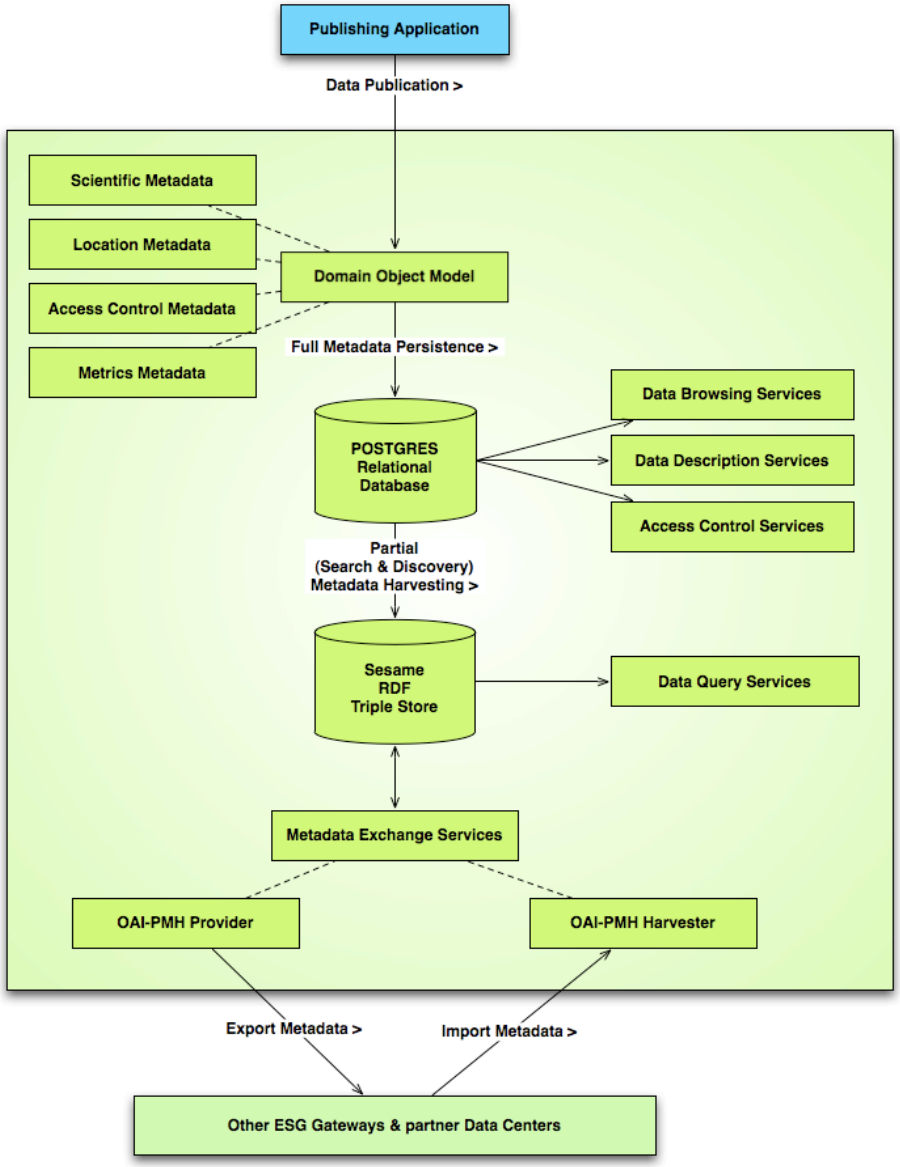


Figure 8. The ESG architecture pipeline for metadata ingestion, usage, and exchange.

4.5 Security

ESG-CET invested a substantial amount of time and resource into investigating, analyzing, and selecting a security strategy along with core technology components. ESG now employs the OpenID framework, a decentralized identity management system, for security. With OpenID's Single-Sign-On (SSO) infrastructure, ESG can manage users at all sites, allowing access only to the permitted set of resources for each user. SSO systems are gaining popularity on the web for applications requiring resources located at different sites because a user can log but once and obtain access to resources housed at any of the remote contributing sites. OpenID was chosen for ESG because it builds on existing Internet technologies such as HTTP, SSL, and Diffie-Hellman, and is lightweight in nature (thus not requiring much additional software to be installed at each Gateway site). Like other web-based login systems, each user is presented with a login screen before accessing protected resources. The user then enters his or her unique, individually assigned OpenID, and is redirected to an ESG-operated Identity Provider. This Identity Provider determines if the user is valid, based upon who the user claims to be and what the user is attempting to access; ultimately, the Identity Provider makes the decision whether to grant access. For supporting single sign-on with non-browser clients, MyProxy Online CA is used to issue user credentials and to provision client machines. MyProxy servers have been deployed on the ESG resources and are primarily used by clients that transfer data and publish meta-data. Where necessary, user attributes are embedded in the issued credentials (which can be used to facilitate authorization based on attributes for non-web environments).

OpenID is a critical system that now ties ESG together through identity management. While Single-Sign-On (SSO) was one of the initial goals of our OpenID evaluation, it is used for many reasons including ease-of-use and installation, a small software footprint, and a proven security model. In particular, ESG leverages the open source OpenID4Java project, a library written in Java that provides an API for using both client and server based functionalities as well as a method for the secure exchange of attributes. OpenID was originally designed to be an open and flexible web-based system that encourages sites to allow users authenticated from arbitrary Identity Providers access to their resources. Potentially, this allows an Identity Provider to be developed that will grant all users permissions to access any resource. For use in ESG, this particular aspect is not acceptable, and so the OpenID4Java project had to be extended to support a few features that were not provided by default.

The first extension to the OpenID4Java project is what we've termed IdP Whitelisting. IdP Whitelisting is a software feature that will only allow authentications from a particular and known list of Identity Providers. The reason for this is that within ESG, it is reasonable for a site to know who authorized the user to access a particular resource, in the interest of security. While OpenID will generally allow authorization from any IdP, the ESG sites have a small acceptable list of providers that must be used. To meet this goal, we have added an API that allows a list of Identity Providers to be allowed during the discovery phase, while potentially eliminating others. In addition, a configuration based reference example has been provided.

The second extension is an API for interfacing with attribute providers. The ESG System has determined that if a user is authenticated by an Identity Provider, attributes will need to be retrieved for the user that can be used at a later time (for example, in performing an attribute based authorization at the resource itself). While OpenID has a method of exchanging attributes, it does not provide a standard API for retrieving attributes. To meet this goal, we have added a flexible API to OpenID4Java that allows a module to be written that can use any method to retrieve attributes. A database module has been

developed and provided as a reference example. This module demonstrates retrieving attributes from a relational database.

Both of these extensions have been added to the version of OpenID4Java used by ESG and are already integrated into the ESG-CET Gateways, which now support the OpenID protocol. However, not all security aspects are web-based, as ESG has clients that perform operations such as data transfer and metadata updates. For these clients, we have put a security system in place that leverages MyProxy.

MyProxy is a credential management system based on X509 Public Key Infrastructure (PKI). Once configured, it allows credentials to be issued to specific users, potentially for a fixed amount of time. This secure method of issuing credentials ensures a user has proper access to a given resource and service. ESG uses MyProxy installations on the Gateway systems and the credentials are issued to non-web based clients (e.g. GridFTP). Extensions to the MyProxy server have been added that allow custom attribute information to be embedded within the issued credentials on a per-user basis (See Figure 9).

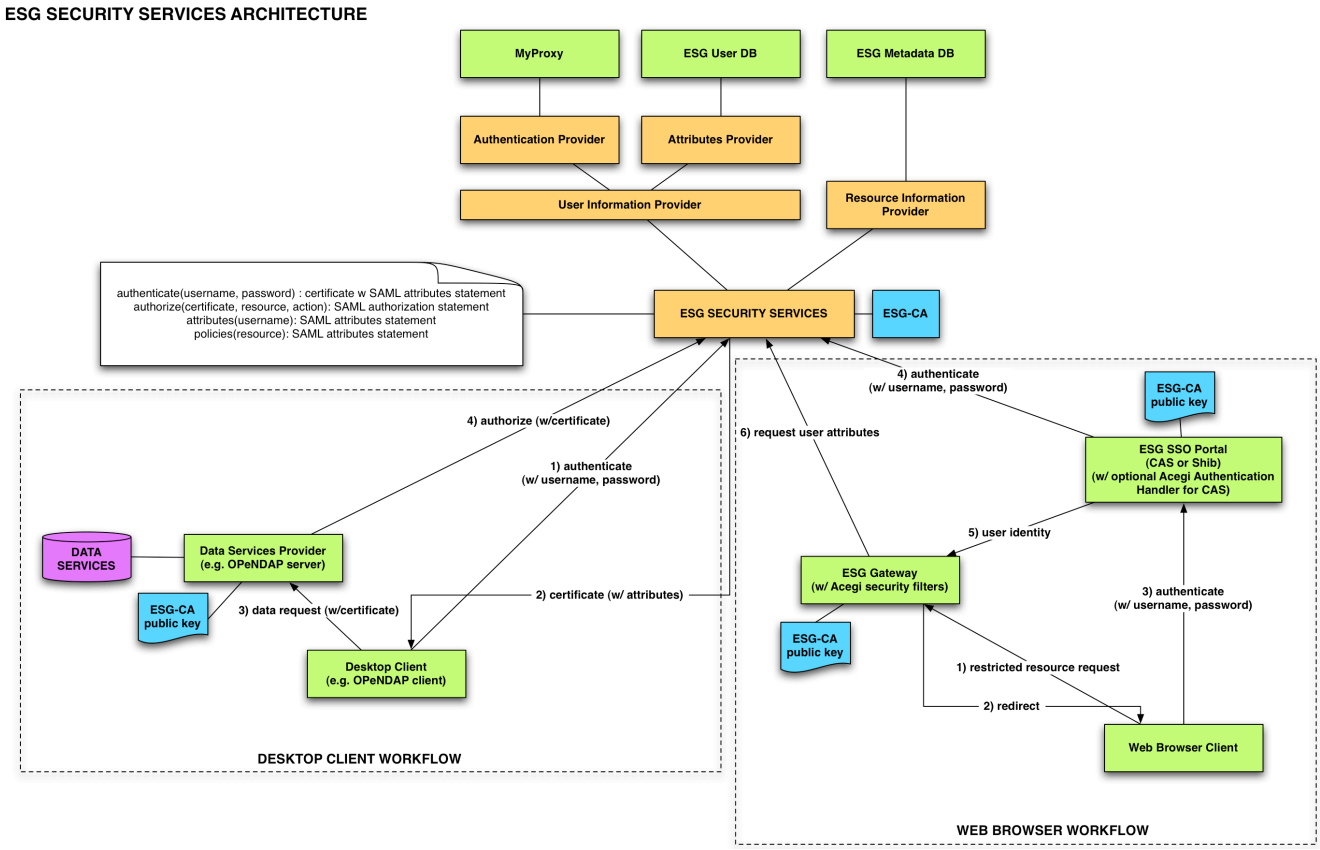


Figure 9. ESG's security services architecture.

4.6 Storage Resource Manager (SRM)

Storage management is one of the most important enabling technologies for large-scale scientific investigations. The ESG project must access data from distributed heterogeneous storage systems, including Mass Storage Systems (MSSs), move data efficiently between storage systems, and manage the content of data in the storage spaces. Modern computational facilities, featuring ever faster computational platforms and increasingly large and complex storage capabilities, produce rapidly growing data volumes. Managing these volumes, especially the extraordinarily large data sets held in

federated distributed centers, requires sophisticated Storage Resource Managers (SRMs), software components whose function is to provide dynamic space allocation and file management on shared storage systems. An SRM calls on data-transport services to bring files into its data space transparently, providing effective sharing of files.

Designed to request multiple files concurrently, thus taking advantage of the available bandwidth, SRMs are based on a common specification that has emerged over time and evolved into an international collaboration. Having an open standard specification that can be used by individual institutions to adapt to their own storage systems has proven to be a remarkable success. SRMs have been used by several facilities, providing uniform interfaces to various storage systems at multiple sites both in the US and in Europe. In ESG, SRMs provide a front-end for various disk-based systems and mass storage systems such as HPSS (High Performance Storage System) at ORNL and Lawrence Berkeley National Laboratory (LBNL) and MSS at NCAR. This arrangement makes it possible for the Gateway site to access all the remote data sources through a uniform interface. Furthermore, the Gateway uses an SRM to manage its disk pool for providing file access to users. When users request files, the Gateway uses the SRM to acquire files from remote systems (if necessary), allocate a quota space for the user, and clean up the space after the user downloads the requested files.

SRM technology has been used successfully to provide uniform access to different Data Nodes, including nodes that keep data on Mass Storage Systems, such as HPSS. SRMs were installed in several nodes, and provide multi-file transport between the nodes and the Gateway based on user's requests. Another product, the DataMover-Lite (DML) has been developed by ESG to provide a portable (PC, Macs, lightweight application that, working in tandem with the Gateway's selection interface, allows a user to easily request a large collection of files for transfer to their portable system, including those on deep storage. In addition to the above functionality, ESG-II adds an emphasis on the use of SRM technology for management of shared disk storage at the Gateway nodes, as well as managing large-scale multi-file transport between nodes, referred to as bulk data movement (BDM).

Since an ESG Gateway's disk is used to stage temporarily files that can be accessed by users, there is a need for managing shared disk on the Gateway. This functionality includes managing space quotas for users, setting lifetimes for the files, and cleaning of the disk space after files are downloaded to the users if the space is needed ("garbage collection"). This functionality is provided by the new recently developed Java-based SRM, called the Berkeley Storage Manager, or BeStMan. BeStMan is also designed to provide file-streaming capability, which allows large transfers to be performed piece-wise in a streaming fashion. Once a request is made, BeStMan brings as many files as will fit into the client's quota space. If some files are already in BeStMan's disk cache (for instance, they were brought in for another client previously), they are made available immediately. The client can now get the available files and "release" them when finished. For every file released BeStMan brings in another file not yet accessed, thus "streaming" the files to the client.

The Multiple File Transport (MFT) service accepts requests for multiple file transfers, and is responsible for reliably delivering the files to various destinations as efficiently as possible. In a typical scenario, a simulation is run at a node, produces a large dataset (10's of terabytes organized as thousands of files) that is stored at that site. At the end of the run, the dataset gets published to its associated Gateway. Next, the dataset or a subset of it needs to be moved reliably to another site in order to be combined with other subsets to form a "core dataset". A core dataset is of interest to a large number of the community members. This MFT task is tedious, and error prone if done manually or with scripts. Providing an automated way of transferring a large number of files reliably and monitoring the transfer progress is an

essential capability for ESG. After the core dataset is generated, it is necessary to replicate (mirror) a subset or all of the core dataset to other sites (perhaps in another part of the world) so the data is closer to the locations where the analysis will be performed. However, the subsets are large enough that they have the same requirements of reliable, recoverable multi-file transfers.

While SRMs are designed to provide reliable and effective multi-file transport, bulk data movement (BDM) is a special case. In practice, one cannot rely on the availability of SRMs at every node site, nor can one rely on having ports open to access the SRM servers because of security limitations. Thus, the bulk data movement has to act as a client program that "pulls" the data on demand. For ESG-II we will design and develop a BDM client that will act as a client version of an SRM. It will be invoked on demand, perform concurrent data transfers reliably (monitor transfers and recover from transient errors), and can be aborted, stopped and restarted on demand. Because BDM replication requests can take a long time to complete, the BDM client must be asynchronous; that is, run as a background task whose status can be checked at any time. The BDM client is planned to become operational within the next year of the project.

4.7 Monitoring

One technology that has contributed significantly to the robustness of the ESG infrastructure is MDS, the Globus Monitoring and Discovery System. MDS monitors the status of components in the distributed system, including GridFTP services, SRMs, the NCAR portal, http data services, the OpenDAP service, and replica catalogs at all the ESG sites. The MDS consists of two components. One is the Index Service collecting status information from providers at each ESG component, including whether a particular service is currently working correctly. The second is the Trigger Service, which takes action based on monitored conditions. In particular, the Trigger Service sends emails to the ESG administrators' mailing list when components fail. This has resulted in much faster recovery of failed services in the distributed ESG infrastructure, resulting in less overall downtime. Prior to the deployment of the monitoring system, it was not uncommon for failures in the infrastructure to be first detected by ESG users, resulting in longer unavailability of services or data, as well as frustration on the part of our users. With the help of the automated monitoring provided by MDS, the ESG team is quickly informed when components fail, allowing the team to quickly restart failed services.

Monitoring operation of the various components of the distributed ESG environment is central to keeping the system operating effectively and ensuring a good user experience. The ESG's monitoring capability is primarily focused on the needs of the ESG staff that maintain the ESG systems, but it also provides useful information to users.

Analysis and experience show that the ESG user community can generally tolerate service outages of reasonable duration, as long as they have adequate information about the situation. Therefore, the ESG team makes an effort to announce planned outages (e.g. preventative maintenance on an ESG system) in advance, and the user-facing gateways include live status information on the various components of the ESG at each participating site, fed by the monitoring system. The ESG team receives more detailed information from the monitoring system via automatic email messages.

The production monitoring system is based on the Globus Monitoring and Discovery System (MDS), which collect information from probes for each of the ESG's component services. When a service is unresponsive, MDS triggers notifications to the ESG team, and appropriate staff undertake diagnostic and corrective action. The systems provide controls to tune the threshold for sending notifications (for

example, brief network interruptions are frequent, and can be perceived as service outages unless the monitoring service checks for multiple successive failures of the probe), and frequency of notifications.

For the next-generation ESG environment, we anticipate that the same basic MDS-based infrastructure will provide us the capabilities we need. Instead, we are focused primarily on the need to extend the breadth and depth of the monitoring applied to each component of the system. An example of broadening the monitoring of the ESG enterprise is to include not just liveness of the various services, but also to monitor resource utilization of the underlying systems. This capability, which we anticipate providing by integrating local monitoring systems like Ganglia or Nagios with MDS, will give managers of ESG sites and the ESG-CET team a better handle on the overall health and stability of the system and help plan for additional resources that may be required to support increasing usage over time. Examples of extending the depth of monitoring would include more sophisticated probes of individual services and multi-site monitoring (executing remote probes from multiple locations) in order to provide more specific and reliable information to the ESG operations staff.

We are currently collaborating with the Metrics working group to understand the extent to which Monitoring and Metrics can share some of the same data collection, aggregation, and reporting infrastructure. While it is not universally the case, for many types of information, the difference is largely the fact that Metrics provides an historical record while Monitoring provides real-time information.

4.8 Metrics

The metrics subsystem has two separate parts: metrics gathering and metrics reporting. Metrics gathering stores User metric data from interactions with the ESG portal into a relational database. Metrics reporting allows System Administrators to analyze how users are using the system. The release scheduled for summer 2009 will include metrics gathering on the Gateway and metrics reporting.

The first release of the ESG Portal has a simple metrics gathering design that will be expanded in future releases. Metrics gathering occurs on both the ESG Web Application and the Data Nodes that make up an ESG Portal. All metric data is stored into a metrics database schema that is part of the ESG Web Application.

Metrics being gathered includes:

- File Downloads – The files that a User downloads.
- User Registrations – The number of Users that have registered with the system.
- User Logins – Tracks the frequency that Users are using the ESG Portal.
- User Searches – Tracks how Users are using the ESG Portal search capabilities.
- User Agents – Tracks which web based clients that are accessing the ESG Portal. User Agents include Web Browsers (Internet Explorer, Safari, Firefox, etc…), Java Clients (Publishing Client), Search Spiders, etc…
- User Clickstreams – Paths that a User takes through the ESG Portal.

There have been many decisions made on the architecture of metrics reporting. The focus for the first release has been on metrics gathering instead of metrics reporting. There are two types of reports that need to be handled by the system: standard reports and ad-hoc reports. Standard reports have been identified as reports that will be run repeatedly by many different Users. Standard reports will be

implemented by integrating with a third party reporting package called JasperReports. JasperReports was chosen because it allows reports to be written quickly due to developer familiarity. JasperReports supports the creation of html, Microsoft Excel, Microsoft Word, and Portable Document Format (pdf) output. The different report output options allow for flexible report delivery options. Reports can be generated at request time through a web page request or at a predetermined time and delivered via an email attachment.

Ad-hoc reports are reports that are created for one-time or sporadic use by a User. Writing SQL by hand and then executing directly against the metrics database schema will create ad-hoc reports. Only specific Users will be granted authorization to perform ad-hoc reports. Ad-hoc reporting is likely to be the main reporting mechanism for the first release.

There are many plans for the future development of the Metrics subsystem. Implementation of a Star Schema for the Metrics data in the database will assist in keeping report generation responsive even with really large tables. More standard reports will be added and made available for Users to run. We will design and implement an automated report mechanism that will deliver reports to Users via email. Finally, we plan to design and implement a Data Node Metrics Software Package for gathering metrics data from a Data Node's LAS, OPeNDAP, and SRM subsystems. The software package will have a web service that will be called by the ESG Web Application that will then download the metrics data and store that data in the metrics database.

4.9 Versioning

Versioning and change notification of the data holdings for ESG-CET came to the forefront as a critical need. In January a working group was established consisting of various ESG-CET stakeholders, including members from the international community, to define the use cases and requirements. The working group identified core use cases for versioning data and notifying users of important changes.

The core use cases for data versioning were identified as:

- updating the files in a dataset;
- triggering a version change;
- retracting a dataset when an error is discovered;
- 'Tagging' the holdings to mark a particular point in time; and
- triggering change notifications.

The core use cases for data users were identified as:

- reviewing changes that have occurred for a dataset;
- subscribing to change notifications;
- retrieving data from older versions; and
- browsing and retrieving data from 'Tags'.

Once the key use cases were identified, more detailed system requirements were derived. Datasets were identified as the central component of versioning system. The key requirements for dataset versioning are:

- any versioning scheme used by the data provider will form the core versioning schema used by ESG-CET; if the data provider does not use a schema of their own, a simple default versioning schema will be used internally by ESG-CET;
- datasets will be the central versioned element;
- data versioning will revolve around versioning of the container dataset. The associated data files will be versioned via their dataset association, not the files themselves;

- only metadata that is directly encoded in the file level metadata will be versioned; and
- dataset version changes will only be triggered by file changes. Any change to a file's checksum should trigger a new version. Adding or removing files from a dataset will also trigger a version change.

While versioning revolves around datasets, it was determined that tracking changes to other metadata elements was important. All metadata objects in the system will have a change log associated with them. The change log will include when the change occurred, who made the change, and a message outlining what the change was. Users will be able to review the logs for other metadata elements in the system.

Data users will be able to register themselves as subscribers to datasets of interest. Once subscribed to a dataset, users will receive e-mail notifications when important events occur. These events can include:

- availability of a new version;
- important changes by the data provider that did not result in a version change; and
- retraction of a dataset due to a critical error

The requirements will be the input source for the next phase of design and implementation of the versioning, notification, and replication aspects of the ESG-CET system.

4.10 Data Replication

Replication of climate data sets was not initially a goal of the Earth System Grid project. Because of the large size of climate data sets, replicating them to multiple sites was considered to be impractical for the first several years of the project. Over time, as the importance of the IPCC data sets stored at PCMDI has increased, several international sites expressed interest in replicating or "mirroring" key portions of the IPCC data sets. Replicating these key data sets has several advantages. Scientists in a particular geographical region would have access to a Figure nearby copy of the data, reducing wide area latencies for data access. Having multiple copies of data sets also provides an increased level of fault tolerance, since data sets are available at other sites even if one site fails or becomes unavailable.

In the past year, the ESG project has worked toward defining the use case for data replication and mirroring with the help of our collaborators at potential mirror sites in the UK and Germany. Initially, we expect that a key subset of the data (the "common core"), which represent approximately 10% of the total data, will be extracted and made available from PCMDI. Mirror sites can replicate this common core or they may construct their own subset of the data for mirroring.

Replication of data sets to a mirror site requires copying the relevant data sets to a Data Node at the mirror site, copying the necessary metadata associated with those data sets to the mirror site's Data Node, and publishing the replicated data sets by making them visible to users of the mirror site's Gateway.

Implementation of the data replication functionality has begun. This work involves the integration of several key ESG components. Once the data sets to be replicated have been identified, the replication service invokes the Bulk Data Movement component to move the data sets reliably. The replication service uses existing ESG metadata API operations to query and replicate the relevant metadata information to the mirroring Data Node. The replication service will use a modified version of the ESG publication client to publish the newly replicated data sets at the mirror site's Gateway, where publication entails making these data sets visible and accessible through the Gateway. Updates to data sets will be identified and propagated to mirror sites by making use of the Versioning functionality. In addition to integrating these existing ESG components, the replication service is responsible for

choosing among available source replicas for the data and metadata and eventually for subscription and notification operations.

Initially, we will use manual notifications to inform mirror sites when new data sets are available or when already-mirrored data sets have failed. In the future, we plan to deploy an automated notification/subscription system. Once a mirror site has replicated a data set, it can register a subscription to that data set. If the Data Node that originally published a data set updates the data set, it issues a notification that the data set has changed, which will automatically trigger notifications to any sites that have subscribed to information about that data set. We hope to use an available open source notification/subscription system, but may implement this functionality ourselves.

Our plans for the remainder of the project with respect to data replication include completing the implementation of the data replication service, deploying it and testing it within ESG. We will also work with collaborating institutions in the UK and Germany, making the data replication software available to them and providing support as they mirror data sets. One of our design goals for the replication service is to provide well-defined interfaces to functionality such as bulk data movement, metadata queries, publishing, etc., that allow our collaborators to use the portions of the ESG replication functionality that are most useful to them without requiring them to deploy the entire ESG software stack. We hope to make it possible for our collaborators to use the ESG replication tools along with their own data movement services or metadata catalogs. We expect that additional mirror sites will also join the collaboration, possibly including sites in Asia and Australia.

4.11 Software Packaging

In preparing the ESG for large-scale production (for CMIP5 and beyond), the deployment and management of the ESG software stack has become a significant issue. We expect the system (particularly the “Data Node” capabilities) to be deployed to several dozen sites around the world. Obviously most of these will be outside the immediate ESG-CET team, and thus outside of our direct control. The ability for people outside the ESG-CET, who are not intimately familiar with the software components involved, to effectively configure, install, and manage the ESG software on their systems, with minimal need for intervention by the ESG-CET team, will be important to ESG’s long-term success and growth.

There are many possible approaches to software packaging and deployment available, with both common and distinctive features. Based on experience with previous versions of the ESG software stack, and observation of packaging and deployment issues in other projects, it seemed important that our approach to software packaging and deployment be able to stand apart from the main operating system software installation in order to better control versioning of software dependencies, and to minimize the potential for adverse interactions with other services that many sites plan to deploy jointly with their ESG installations. Also, it seems useful and desirable to use a tool that supports basic ideas of software management, such as installations, updates, and removal, use of remote repositories, etc.

To fill this need, we are currently evaluating the Pacman package management tool used by the Open Science Grid’s (OSG) Virtual Data Toolkit (VDT). Pacman has been widely used and strongly influenced by grid projects such as International Virtual Data Grid Laboratory (iVDGL), TeraGrid, and especially the OSG and VDT. The VDT team has become one of the major contributors of Pacman software. The ability to leverage the existing investment in packaging and testing of the VDT software stack is a major consideration in pursuing Pacman. The ESG software stack consists of a combination of project- and climate-focused packages together with more widely used grid and other infrastructure.

There is a reasonable degree of overlap between the infrastructure required by ESG and that provided in the VDT. In many cases, the specific versions of packages in VDT differ from those currently targeted by ESG development, but based on discussions with the OSG project, we do not believe this is a major issue should we decide to commit to Pacman. Another advantage of the Pacman/VDT environment and collaboration with OSG would be easier access to the NSF Middleware Infrastructure (NMI) Build and Test Lab led by the University of Wisconsin. The Build and Test Lab provides testing facilities for distributed software systems on a wide range of operating systems, which would be valuable in hardening both the ESG software itself, as well as the configuration/build/install process. Since VDT is tested using this facility, extending such tests to cover the ESG's needs would be easier than building them from scratch for another testing environment.

Having evaluated the basic features and capabilities of Pacman, we are currently in the process of creating a package for the ESG Publishing tool, which is central to the Data Node software stack, along with its dependencies. Experience thus far shows that it is relatively straightforward to create Pacman packages that allow non-root users to install software in locations separate from the main system software tree (i.e. /bin, /lib, /usr/local/, etc.). We expect to complete this evaluation fairly soon and come to a final decision as to whether to adopt Pacman and VDT.

Appendix A Collaborative Interactions

To effectively build an infrastructure capable of dealing with petascale data management and analysis, we have established connections with other funded DOE Office of Science SciDAC projects, programs, and agencies at various meetings and workshops. This section describes many of the substantive interactions we've had during the first half of the project.

A.1 October 1, 2008 through March 31, 2009

A.1.1 External Support of Grid-Enabled OPeNDAP

Prof. Giovanni Aloisio and Dr. Sandro Fiore of Scientific Computing & Operations, CMCC-Euromediterranean Centre for Climate Change, who are working on the Climate-G project, also are collaborating with the Data and Transport ESG-CET team. OPeNDAP efforts at NCAR, focusing on data infrastructure and GridFTP developments, are testing and incorporating into their system.

Dr. Monique Petitdiddier, Dr. Horst Schwichtenberg, Dr. Wim Som de Cerff (and collaborators) from the EGEE - Enabling Grids for E-sciencE project have been collaborating on numerous aspect of data transport within Grid systems with the ESG-CET team.

A.1.2 University of Utah Visit to Discuss VisTrails, ESG, and CDAT Collaboration

As part of the SciDAC VACET collaboration effort, ESG is working closely with the University of Utah to integrate one of the ESG analysis package (i.e., The Climate Data Analysis Tools (CDAT)) into Vistrails (Figure 10). (VisTrails is an emerging paradigm for capturing complex analysis processes at various levels of details and provides provenance information necessary for reproducible results in a shared environment.)

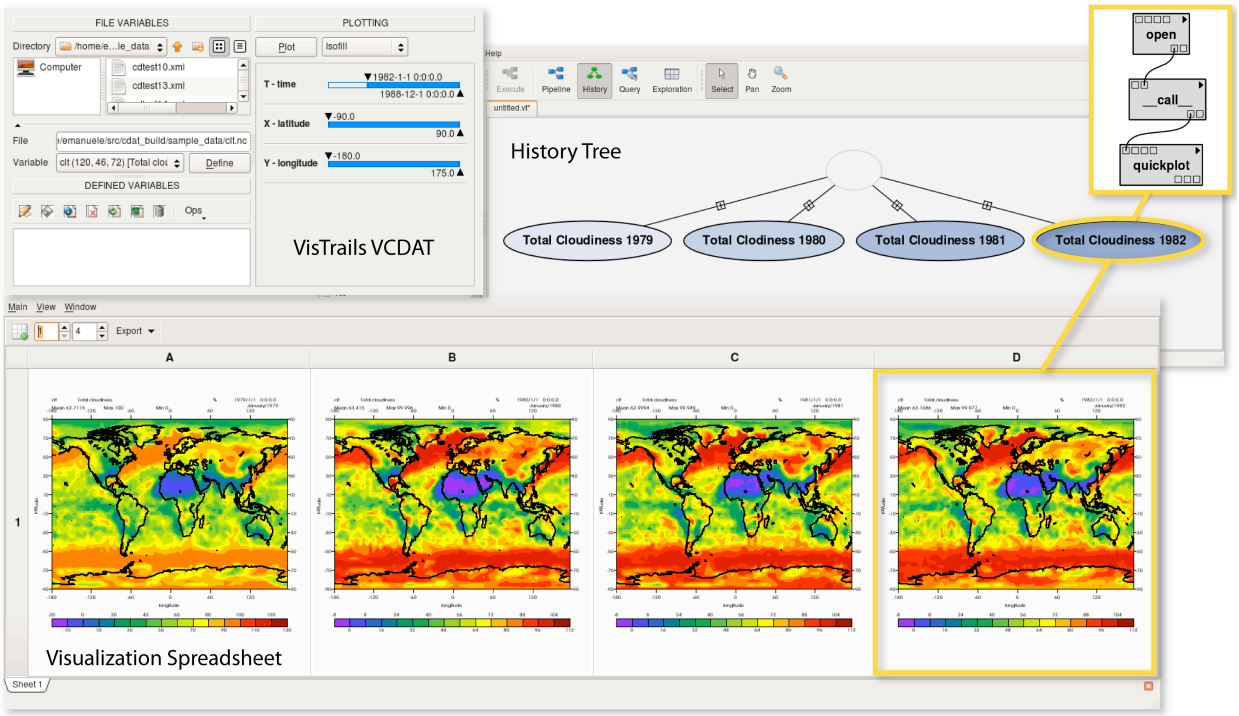


Figure 10. CDAT VisTrails package usage. Users interact with the VisTrails VCDAT window to manipulate the climate datasets and visualize them in the Visualization Spreadsheet. Provenance is automatically captured as a tree, where each node corresponds to a workflow that was also automatically generated based on the actions performed on the Vistrails VCDAT window.

A.1.3 Collaboration with the SciDAC VACET Visualization Team

The ViSUS 2.0 framework has been released as part of latest CDAT (i.e., version 5.0). This result follows from a substantial amount of software engineering effort over the past 18 months. This result is significant because all CDAT users now have access to a solid set of fundamental 3D visualization capabilities, of which will be eventually accessed via ESG (Figure 11 and 13).

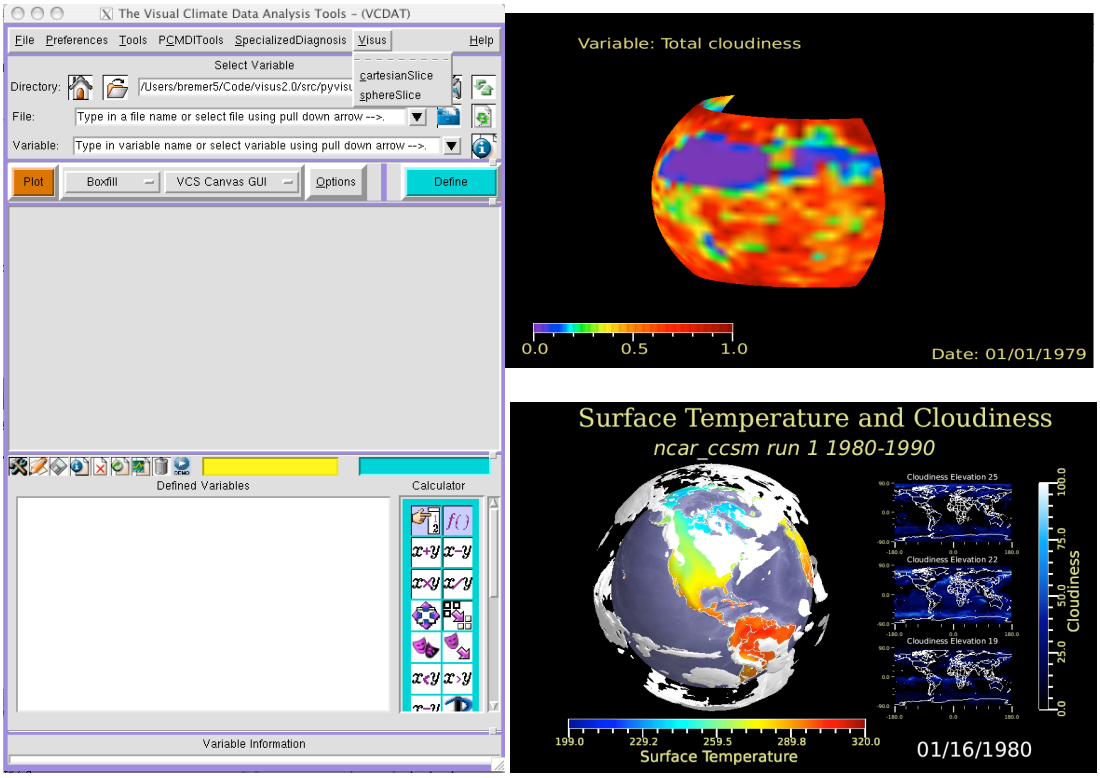


Figure 11. Custom VCDAT interface showing the new ViSUS capabilities. Two images showing the partial mapping of cloudiness and a scene showing 3D iso-contouring, slicing, and multiple 2D plots.

A.1.4 Global Organization for Earth System Science Portal (GO-ESSP) Workshop

Since its conception in 2002, ESG-CET leaders Steve Hankin (2008 workshop host), Don Middleton, and Dean N. Williams (2006 workshop host) have participated in the GO-ESSP steering committee. The GO-ESSP workshop focuses on facilitating the organization and implementation of an infrastructure for full-data sharing among a consortium spanning continents, countries, and intergovernmental agencies.

A.1.5 Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)

NOAA/PMEL (Steve Hankin, ESG co-PI) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium [http://hycom.rsmas.miami.edu/]. The HyCOM Consortium has developed a high resolution (1/12 degree) operational, global ocean modeling capability under cooperative US Navy and NOAA funding. The HyCOM model presents unique technical challenges, through the complicated coordinate system that it employs and its large data volumes, but the needs of HYCOM overlap in many respects with the ocean components of the climate models to be utilized in IPCC AR5. There is a significant and productive two-way technology transfer of technical capabilities developed in support of ESG and technical capabilities developed in support of HyCOM

A.1.6 NOAA Geophysical Fluid Dynamics Laboratory

The NOAA GFDL Fluid Dynamics Laboratory is an active contributor to AR5 and an active participant in the ESG SciDac. V. Balaji [Head, GFDL Modeling Systems Group] is a frequent participant and active contributor in ESG telcons and meetings leading to a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin, ESG co-PI) shares an MOU with GFDL for the

development of the Laboratory's data portal, also leading to an active two-way technology transfer between NOAA and ESG.

A.1.7 NOAA Office of Climate Observations (OCO)

PMEL is the developer of the ocean Observing System Monitoring Center (OSMC) on behalf of NOAA/OCO and manages interactive access to the international Surface Ocean Carbon ATlas (SOCAT) for quality control analysis. Through the PMEL membership in the ESG SciDAC a number of useful collaborative benefits are being explored and are likely to be realized in time for IPCC/AR5 work. OSMC and SOCAT are both sources of integrated ocean-climate observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be helping to bring these collections of observations into the ESG framework for the benefit of IPCC scientists and others.

A.1.8 Unidata and the Climate and Forecast Conventions (CF)

Several ESG members play key roles in the development of the CF conventions – the emerging standard for climate model outputs stored in netCDF. ESG is forging a strong collaborative relationship with Unidata, the development organization for netCDF.

A.1.9 US Integrated Ocean Observing System (IOOS)

PMEL is a member of the US Integrated Ocean Observing System (IOOS) Integrated Products Team (IPT). IOOS is a potential source of integrated ocean observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be collaborating with IOOS to locate climate-relevant US coastal observations and bring them into the IPCC framework.

A.2 April 1, 2008 through September 30, 2008

A.2.1 University of Utah Visit to Discuss VisTrails, ESG, and CDAT Collaboration

As part of the SciDAC VACET collaboration effort, Dean Williams visited the Scientific Computing and Imaging (SCI) Institute at the University of Utah to discuss the possibilities of VisTrails in the ESG framework. (VisTrails is an emerging paradigm for capturing complex analysis processes at various levels of details and provides provenance information necessary for reproducible results in a shared environment.) The LLNL group was able to install VisTrails and implemented a small subset of CDAT commands callable from within the VisTrail visual environment. Although CDAT has been developed with climate applications in mind, most of the concepts of the system can be used to develop a more general data analysis tool by merging the functionalities of CDAT and VisTrails (Figure).

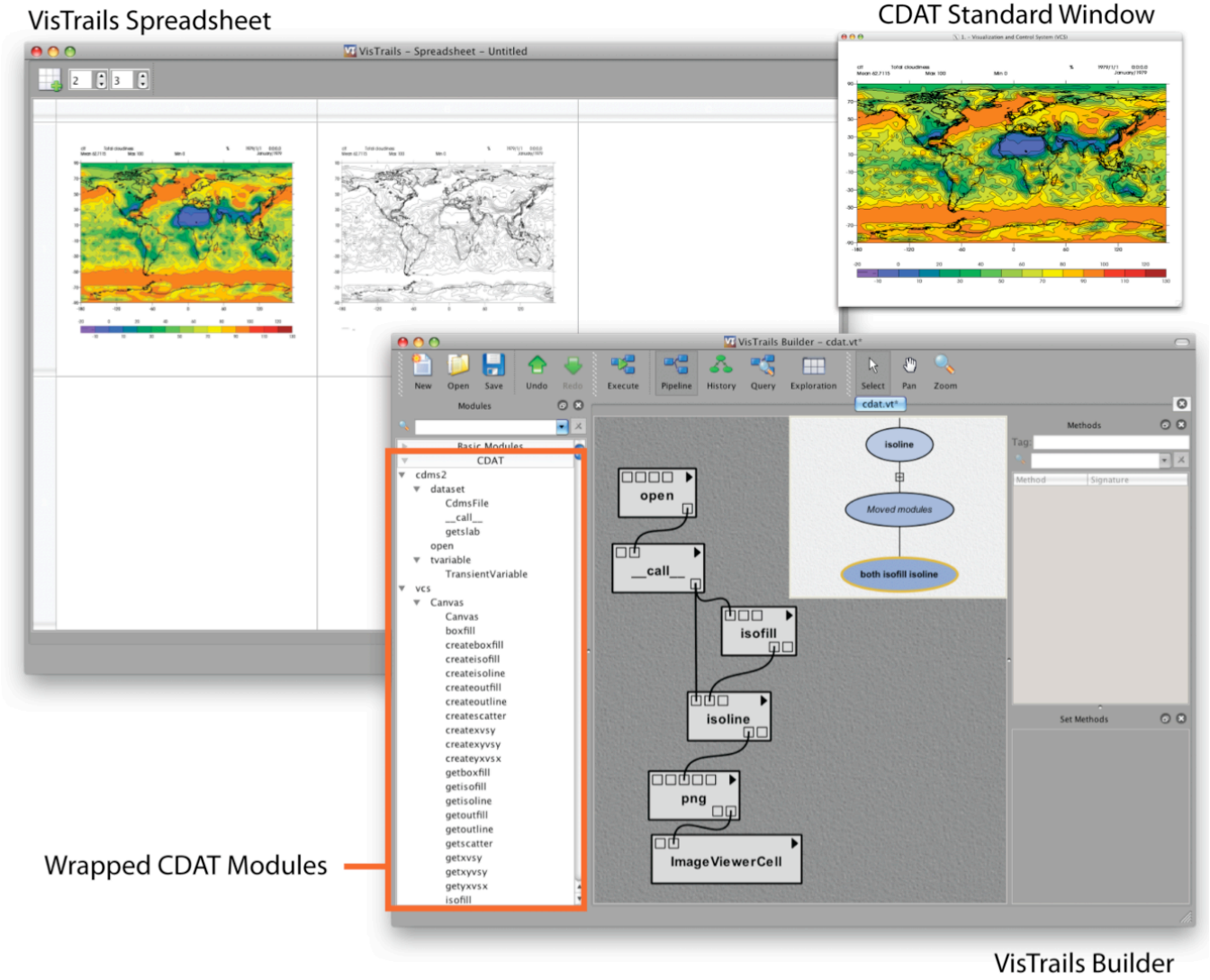


Figure 12. Example of CDAT workflow built inside VisTrails.

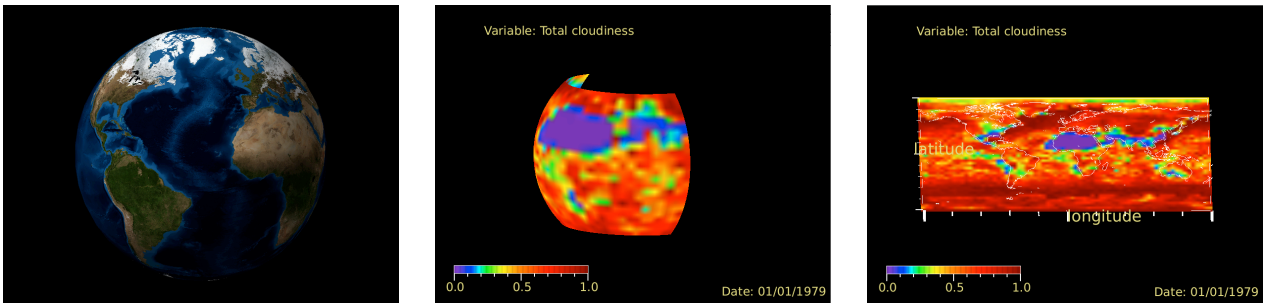
A.2.2 Collaboration with the SciDAC VACET Visualization Team

During the 2008 SciDAC conference the VACET team demonstrated the new 3D visualization system add-on to the 5.0 beta release of the CDAT package to John Drake and Phil Jones. Both were enthusiastic about the 3D capabilities (Figure 13), their tight integration into the existing pipeline, and especially the flexible annotations offered by the new system. John and Phil re-iterated their concerns about the high spatial resolutions of upcoming simulations and the inability of existing tools to adequately handle these large data sets. All parties agreed that the ViSUS package and its ability to process tera-scale datasets in a streaming and/or out-of-core fashion is an important and welcome addition to the CDAT package.

Going forward, John and Phil suggested two climate-specific extensions for the current tool chain: Mapped z-elevations and mapped logical grids. Mapped z-elevations are important for climate visualization since most common climate simulation codes operate on slices of constant air pressure rather than constant elevation. While both measurements are related, they are not identical and their

mapping must be taken into account for an accurate visualization. The current version of the ViSUS system already has the capability to display height fields, as mapped 2D planes, and this capability will be extended to support general elevation mappings. While older simulation codes used polar grids with discontinuities at both geographic poles, newer packages use warped polar grids that move the singularities to less significant locations (e.g., the middle of Canada), or more generally mapped grids.

While these structures logically remain regular grids geometrically they can be mapped almost arbitrarily onto the earth depending on the specific needs of a given simulation. Multiple options to support mapped grids in the ViSUS system were discussed with higher dimensional coordinates being the clear favorite. This concept entails storing the geometric locations of samples along with the standard physical variables (pressure, temperature, etc.). Since the grid logically remains regular, the "enhanced" dataset can easily be handle by the existing streaming infrastructure. The mapping can now be seen as a special case of multi-modal visualization where, for example, temperature is one field and x,y,z, positions are a second field. Custom rendering code will then combine both fields to create a mapped temperature field.



(a) Earth display using satellite imaging. (b) Partial mapping of cloudiness. (c) Continents rendered using new poly-line display capability.

Figure 13: Example 3D Capabilities Provided by the VACET Team

One of the first new features that the VACET team will introduce is the potential management of multiple meshes with different domains and resolutions that may be incompatible. In particular, the VACET team will work on the visualization of atmospheric and ocean data that are generated on independent domains.

A.2.3 Collaborations with the Southern California Earthquake Center (SCEC)

Ann Chervenak initiated discussions on possible collaborations regarding metadata schemas and catalogs with the ESG-CET and the Southern California Earthquake Center (SCEC) project. She also led phone meetings between the two projects. Participants from SCEC included Thomas Jordan (PI), Phil Maechling, David Okawa and Tran Huynh – all from the University of Southern California's Information Sciences Institute (USC/ISI).

A.2.4 Scientific Data Management (SDM) Center for Enabling Technology (SciDAC CET)

Based on their experience with DataMover-Lite, Arie Shoshani and his team developed a new client version of an SRM (called "SRM-Lite") in order to move files to and from sites that have highly secure systems by using one-time-passwords. This tool is similar in design to DML, but runs as a client version of SRM to pull/push files from behind a firewall. Unlike DML, SRM-lite cannot use a GUI remotely, and therefore has only a command-line interface. For example, Chi-Fan Shih of NCAR's Data Support Section used this tool to move ESG data from a disk system inside the firewall to one outside the

firewall. SRM-Lite, based on high-performance scp (hpn-scp) for higher throughput, is being used successfully to move the large volumes of data.

A.2.5 Center for Enabling Distributed Petascale Science (CEDPS) (SciDAC CET)

Collecting statistics on data movement and storage usage is important to ESG. In order to collect such data from BeStMan, Arie Shoshani's team collaborated with members of the Center for Enabling Distributed Petascale Science (CEDPS) project to extract such data automatically from the SRM logs. They have worked with Dan Gunter and Keith Beattie at LBNL for CEDPS troubleshooting to collect logging information into DB, display with NetLogger and web-frontend. They also plan to use their statistics collection tools to automatically generate summary statistics and help with troubleshooting.

ESG also collaborates with CEDPS in the data transfer area, working to define new requirements for GridFTP and for data replication services.

A.2.6 SBIR CDAT-MODAVE – Collaboration with Tech-X

Dean Williams is working with Alex Pletzer (from Tech-X) and V. Balaji (from the Geophysical Fluid Dynamics Laboratory, GFDL) on the Mosaic Data Analysis and Visualization Extension (MoDAVE). An SBIR-funded project, MoDAVE started in July 2008 with the aim of extending CDAT to handle the large multi-block datasets that are increasingly becoming the norm, as atmospheric models move away from longitude/latitude based grids (Figure 14). Since its inception, MoDAVE has focused on developing a cubed-sphere netCDF reader using CDAT's Climate Data Management System 2 (cdms2) object to infer the connectivity between mosaic tiles. Knowledge of the tile connectivity is necessary in order to accurately visualize cell-centered data on the cubed-sphere (finite-volume-based discretization leads to cell-centered scalar fields). A general API (mvViz3d) was defined to allow multiple 3D visualization engines to be plugged into CDAT. At present mvViz3d supports two implementations: one based on the Visualization Toolkit (VTK) and the other on VisIt (www.llnl.gov/visit/); both can be used to render 3D iso-surfaces on the sphere. In addition, we will continue to explore ways to parallelize data analysis and visualization so as to be able to handle high-resolution (< 10km) datasets. CDAT is one of several standard analysis tools that must operate on ESG's "Product Services" back-end.

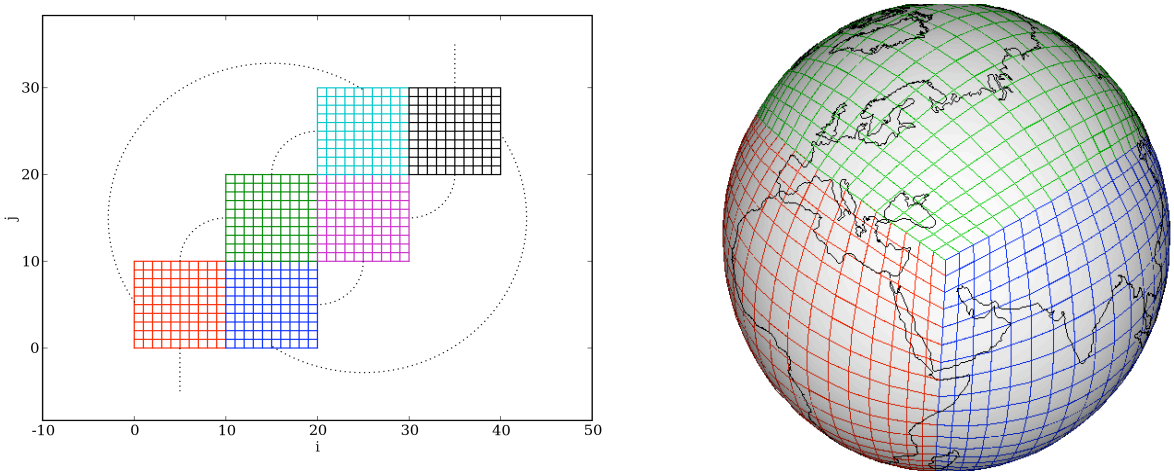


Figure 14: Example of a Mosaic Grid: Cubed Sphere Connectivity

A.2.7 Global Organization for Earth System Science Portals (GO-ESSP) Workshop

Steve Hankin (this year's workshop host), Don Middleton, and Dean N. Williams are three of seven GO-ESSP steering committee members who coordinated the seventh annual GO-ESSP workshop held September 16 – 19 at the Seattle Washington Public Library. In addition, Steve, Don, and Dean chaired six of the seven workshop sessions. The GO-ESSP workshop focuses on facilitating the organization and implementation of an infrastructure for full data sharing among a consortium spanning continents, countries, and intergovernmental agencies. All ESG-CET testbed partners (i.e., LLNL, NCAR, GFDL, BADC, DKRZ, and the University of Tokyo) were present. The workshop, in part, covered testbed security concerns and addressed issues of collaboration. By 2011, this organization envisions allowing users open access to petabytes of multi-model generated data, as well as in-situ, satellite, biogeochemistry, and ecosystems data.

A.2.8 Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)

NOAA/PMEL (Steve Hankin) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium [http://hycom.rsmas.miami.edu/]. The HyCOM Consortium, which has developed a high-resolution (1/12 degree) operational, global ocean modeling capability under cooperative US Navy and NOAA funding. The HyCOM model presents unique technical challenges, such as the complicated coordinate system that it employs and its large data volumes, but the needs of HYCOM overlap in many respects with the ocean components of the climate models to be utilized in CMIP5 (IPCC AR5). A significant and productive two-way technology transfer has been developed in support of ESG-CET, and of HyCOM in particular.

A.2.9 NOAA Geophysical Fluid Dynamics Laboratory

The NOAA GFDL Fluid Dynamics Laboratory is an active contributor to CMIP5 and an active participant in the ESG-CET. V. Balaji [Head, GFDL Modeling Systems Group] is a frequent participant and active contributor in ESG-CET telcons and meetings leading to a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin) shares a Memorandum of Understanding (MOU) with GFDL for the development of the Laboratory's data portal, also leading to an active two-way technology transfer between NOAA and ESG-CET.

4.11.1 NOAA Integrated Ocean Observing System (IOOS) and NOAA Office of Climate *Observations (OCO)*

NOAA has been designated as the lead agency in the development of the U.S Integrated Ocean Observing System (IOOS). PMEL is a member of the IOOS Integrated Products Team (IPT) and is also the developer of the ocean Observing System Monitoring Center (OSMC) on behalf of NOAA/OCO. Through the PMEL membership in the ESG-CET a number of useful collaborations benefits are being explored and are likely to be realized in time for the CMIP5 (IPCC AR5) work. IOOS and OSMC are both sources of integrated ocean observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be helping to bring these collections of observations into the ESG-CET framework for the benefit of IPCC scientists and others.

A.2.10 NASA JPL visit to LLNL to discuss ESG-CET and Satellite Data

Dan Crichton and Amy Braverman from NASA's Jet Propulsion Laboratory (JPL) are actively pursuing initiatives in climate modeling, and so are interested in establishing an active relationship with ESG. Their recent visit (along with other JPL scientists) to LLNL explored how to enable the climate

community to gain access to JPL satellite data, and allow modelers at JPL to gain access ESG data. JPL is now engaged in a pilot activity that demonstrates access between JPL and LLNL and, if successful, will encourage deeper collaborations between NASA and ESG-CET.

A.2.11 External Support of Grid-Enabled OPeNDAP

Karen Schichardt and Jeff Daily of PNNL, who are working with Dr. David Randall (CSU) on an NSF STC for extreme weather modeling, also are collaborating with the Data and Transport ESG-CET team. OPeNDAP efforts at NCAR, security infrastructure, and GridFTP protocol developments now are helping this group integrate their efforts into the ESG-CET environment.

A.3 October 1, 2007 through March 31, 2008

A.3.1 Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)

NOAA/PMEL (Steve Hankin, ESG co-PI) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium [http://hycom.rsmas.miami.edu/]. The HyCOM Consortium has developed a high- resolution (1/12 degree) operational, global ocean modeling capability under cooperative US Navy and NOAA funding. The HyCOM model presents unique technical challenges, owing to its complicated coordinate system and large data volumes, but the needs of HYCOM overlap in many respects with those of the ocean components of the IPCC AR5 climate models. There is thus a significant and productive two-way transfer of technical capabilities developed in support of ESG and HyCOM.

A.3.2 North America Climate Regional Climate Change Project

The ESG-CET collaboration has worked towards enabling support, within the current ESG operational system, for publishing and distributing NARCCAP (North America Climate Regional Climate Change Project) data. An extensive data management plan was developed that involves distributed data access from the ESG portal at NCAR to data resources stored at both NCAR and PCMDI. The existing user registration system was extended to allow a separate community of NARCCAP users vetted by specific administrators, and the first test users were approved for access.

A.3.3 NOAA Geophysical Fluid Dynamics Laboratory

The NOAA GFDL Fluid Dynamics Laboratory is an active contributor to AR5 and an active participant in the ESG SciDac. V. Balaji [Head, GFDL Modeling Systems Group] is a frequent participant and active contributor in ESG teleconferences and meetings, resulting in a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin, ESG co-PI) shares an MOU with GFDL for the development of the Laboratory's data portal, thereby also implementing an active two-way technology transfer between NOAA and ESG.

A.3.4 Global Organization for Earth System Science Portal (GO-ESSP)

The GO-ESSP is a collaboration designed to develop a new generation of software infrastructure that will provide distributed access to observed and simulated data from the climate and weather communities. Of the seven members of the GO-ESSP steering committee, three are members of the ESG-CET team: Steve Hankin, Don Middleton, and Dean N. Williams.

A.3.5 Earth System Curator (ESC)

The ESG-CET and the Earth System Curator (ESC) are working together to develop prototype ontology, user interface, and relational databases to include additional information on the model configurations that produce datasets. ESG-CET team members Luca Cinquini and Don Middleton (as a co-PI) are working closely with this group. Other ESG-CET team members may be involved as work progresses.

A.3.6 Scientific Data Management (SDM) Center for Enabling Technology (SciDAC CS CET)

Based on the experience with DataMover-Lite, a new client version of an SRM (known as "SRM-Lite") was developed in order to invoke it to move files to and from sites that have one-time-password (OTP) security or other highly secure systems. SRM-Lite is similar in design to DML, but has only a command-line interface. We plan to use this tool in a workflow system that the SDM center uses, called Kepler, as well as other application projects.

A.3.7 VACET: VisTrails

VisTrails is a new scientific workflow management system. The LLNL team is currently working to modify the Climate Data Analysis Tools (CDAT) XML output to allow the use of VisTrails. If successful, this workflow may be implemented in the next generation of ESG architecture.

A.3.8 Institute for Ultrascale Visualization

Jian Huang (U. Tennessee), a member of the SciDAC Institute for Ultrascale Visualization, is developing tools for web-based collaborative visualization of climate data, with initial application to data from the ORNL C-LAMP portal. ORNL plans to deploy Huang's visualization server on the ORNL C-LAMP computer, and will work towards integrating it as a "product server" in the next-generation ESG architecture. Huang has been invited to present a paper on this work at the May 2008 International Symposium on Collaborative Technologies and Systems (CTS 2008).

A.4 April 1, 2007 through September 30, 2007

A.4.1 North American Regional Climate Change Assessment Program (NARCCAP)

The ESG-CET collaboration has worked towards enabling support, within the current ESG operational system, for publishing and distributing NARCCAP (North America Climate Regional Climate Change Project) data. An extensive data management plan was developed that involves distributed data access from the ESG portal at NCAR to data resources stored both at NCAR and PCMDI. The existing user registration system was extended to allow a separate community of NARCCAP users vetted by specific administrators, and the first test users were approved for access.

A.4.2 GO-ESSP Collaboration: Semantic Technologies

During the past few months, considerable effort was spent in investigating the use of emerging semantic technologies (RDF, OWL, Sesame) to develop the next generation of ESG-CET services for search and discovery of scientific data. Prototype search services and interfaces were set up against the current IPCC, CCSM and PCM metadata holdings in order to test the performance, flexibility, and scalability of this approach. Although the first results in this area are encouraging, work is still underway.

More recently, discussions have taken place with the Earth System Curator (ESC) collaboration, which has decided to leverage this prototype ESG-CET infrastructure to provide powerful detailed search capabilities for climate models and their components, as described by the extensive ESC metadata

schema. The plan is for ESC to reuse the existing ESG-CET semantic service and persistence layers, collaborating to extend the current ESG-CET ontology with additional classes and properties, while at the same time adding custom functionality for compatibility checking among model components. A meeting will be held at GFDL in mid-October 2007 to assess progress and to plan for the next phases of the collaboration between the two projects.

A.4.3 IO Strategies and Data Services for Petascale Data Sets from a Global Cloud Resolving Mode Collaboration

The ESG executive committee has met with Karen Schuchardt (the SAP PI on Global Cloud Resolving Models) on numerous occasions, outlining the strategy for working together as a team. More recently at the Climate Change Prediction Program (CCPP) conference in Indianapolis, Karen and Dean discussed working more closely at the PI level. The general agreement is to include Karen, once a month, on ESG executive committee meetings (starting in October). This will keep her abreast of ESG activities and help ESG leverage work completed by her team. We also discussed pairing members of her group with working groups already established in ESG: the Metadata Work Group (i.e., working with Bob and Luca on metadata schemas, RDF, etc.), and the User Interface Working Group (i.e., working with Jens and other doing ESG user interface development). Also planned is providing help for the LLNL team to extend CDAT to support a geodesic grid, which also involves Geophysical Fluid Dynamics Laboratory (GFDL) gridspec work. (The results of the gridspec effort, led by V. Balaji at GFDL, will be implemented in the netCDF Climate and Forecast (CF) convention.) In addition, LLNL team members will also discuss the Climate Model Output Rewriter (CMOR) and how to improve processing data for model intercomparisons such as CMIP3 (IPCC AR4).

A.4.4 Atmospheric Radiation Measurement (ARM) Collaboration

The team at Argonne has started collaborating with Environment Science Division at ANL, specifically to work with scientists at Climate Research Station on the Data Domain to Model Domain Conversion Package (DMCP) (see URL: http://www.atmos.anl.gov/DMCP/). This recently initiated effort has been exploring ways to publish subsets of ARM data with mechanisms to support useful parameter-based server-side processing of data. The collaboration also will investigate options to allow publishing the resulting data as an independent dataset.

A test installation of Live Access Server (LAS) has been set up and work is ongoing to evaluate the upload, visualization and processing of a sample subset of ARM data. The results from the evaluation of the prototype will be used in the design and implementation of server-side processing on ESG systems. (See section 2.6.)

A.4.5 Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)

NOAA/PMEL (Steve Hankin, ESG co-PI) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium (see URL: http://hycom.rsmas.miami.edu/). The HyCOM Consortium is developing a high-resolution (1/12 degree) operational global ocean modeling capability under cooperative US Navy and NOAA funding. The HyCOM model presents unique technical challenges, through the complicated vertical coordinate system that it employs, but its needs overlap in many respects with the ocean components of the climate models to be utilized in CMIP4 (IPCC AR5). There is a significant and productive two-way transfer of technical capabilities developed in support of ESG and of HyCOM. (See Figure 15, showing the HyCOM model intecomparison.)

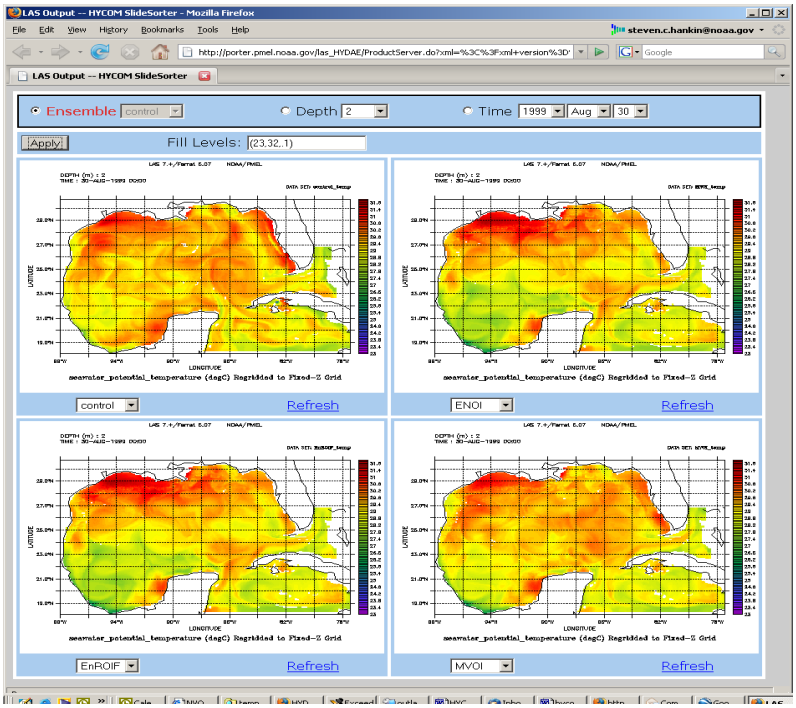


Figure 15. LAS slide sorter output showing HyCOM model intercomparison.

A.4.6 NOAA Geophysical Fluid Dynamics Laboratory

The NOAA Geophysical Fluid Dynamics Laboratory (GFDL) is an active contributor to CMIP4 (IPCC AR5) and an active participant in the ESG-CET. V. Balaji (Head, Modeling Systems Group at GFDL) is a frequent participant and active contributor in ESG telcons and meetings, resulting in a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin, ESG co-PI) shares an MOU with GFDL for the development of the Laboratory's data portal, also affecting an active two-way technology transfer between NOAA and ESG.

A.4.7 Scientific Data Management (SDM) Center for Enabling Technology (SciDAC CS CET)

Similar to the DataMover-Lite (DML) client component in ESG, the SDM center has identified a need for moving files to and from sites that have one-time-password (OTP) security or other highly secure systems. The intention is to have an SRM client program at the secure sites that communicate commands and data through SSH. The SDM center has developed a prototype version of this client program, called SRM-Lite, and plans to use this technology for a combustion project in the near future.

A.4.8 VACET Collaboration: VisTrails

VisTrails is a new scientific workflow management system. While originally (and solely) developed by researchers at the University of Utah to provide support for data exploration and visualization, VisTrails now is being applied to climate data analysis and visualization as part of the SciDAC-2 Visualization and Analytics Center for Enabling Technology (VACET) collaboration. Figure 16 depicts the use of the visual workflow interface to connect CDAT module boxes to perform calculations and a related plot.

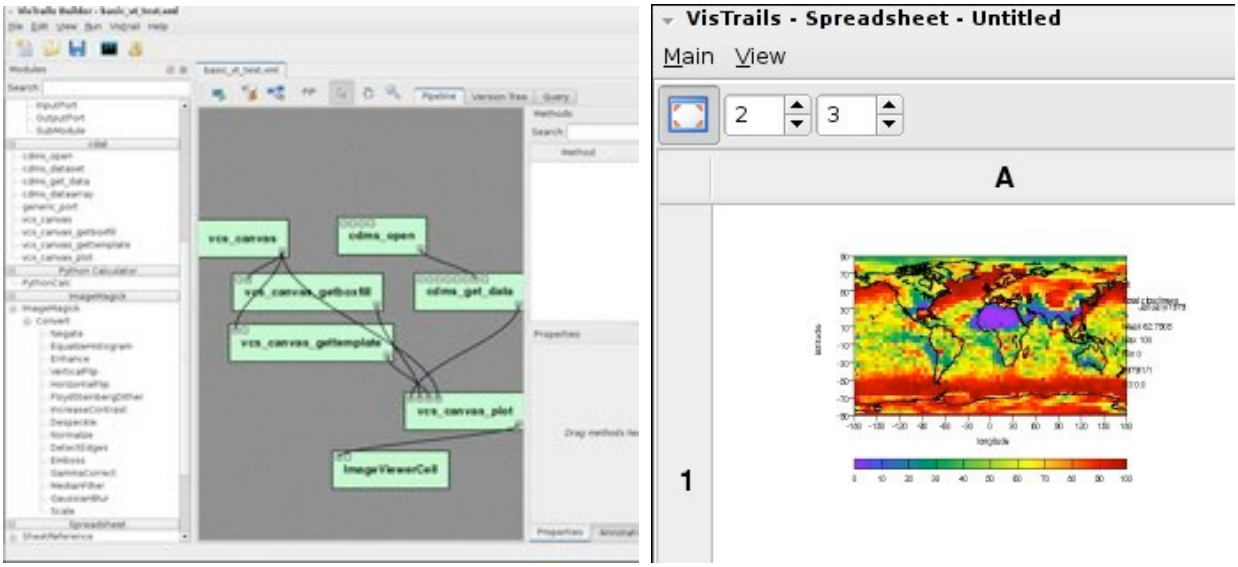


Figure 16. The result of the CDAT run viewed in VisTrails showing results in a spreadsheet application.

Work on this new GUI application interface for climate data analysis and exploration continues in collaboration with the VACET team.

A.4.9 VACET Collaboration: 3D Visualization

In its collaboration with VACET, the ESG-CET team has worked to produce several compelling, high-quality, 3D images that will be reproducible by any scientists who have access to ESG-CET's computational resources to do ground-breaking 3D visualization and computing. Initially, these images would lend themselves to the creation of "glitzy" movies used for general public consumption. In the future, we aim for scientists to produce these images in pursuit of understanding key climate science questions. The visualization appearing below shows surface temperature, atmospheric temperature, and sea ice and cloud coverage on an elevated Earth model. In this example, the data (e.g., surface temperature) represent the combined average influence of an ensemble of all the climate models that are available in the CMIP3 (IPCC AR4) data archive. The animation over time shows an upward climate temperature trend, indicative of global warming. This visualization/animation example was computed on 200 processors in about 15 minutes using custom visualization software that will be integrated into climate analysis tools (Figure 17).

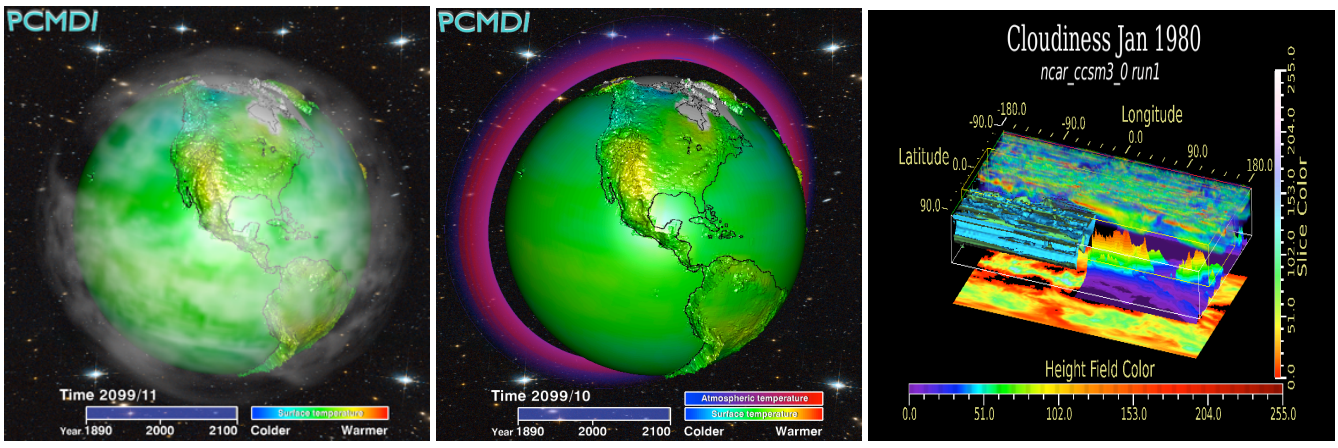


Figure 17. Working with VACET developers to make 3D graphics accessible to the climate community.

A.5 October 1, 2006 through March 31, 2007

A.5.1 A Scalable and Extensible Earth System Model for Climate Change Science (SciDAC Climate Application)

Collaborating with the SciDAC science application project *A Scalable and Extensible Earth System Model for Climate Change Science* (SEESM), ESG-CET has agreed to address the challenges of enabling data management, discovery, access, and advanced data analysis for their community. In constant contact with the SEESM PI, John Drake, and co-PI, Phil Jones, ESG-CET will act as the liaison for this group with other SciDAC projects (such as VACET and IUSV).

A.5.2 Intergovernmental Panel on Climate Change Fifth Assessment Report (IPCC AR5, United Nations Environment Program)

Dean Williams serves as the liaison between the current AR4 Working Group I committee leaders, led by Jerry Meehl, and the future AR5 Working Group I committee, to be led by Ron Stouffer. In October, Dean met with Ron and others to discuss proposed scenarios/experiments for AR5 and to work out strict deadlines for the project.

A.5.3 Computational Climate End Station (ORNL Leadership Computing Facility)

David Bernholdt (ORNL) serves as liaison with the Computational Climate End Station at the Leadership Computing Facility at ORNL. He is in regular contact with the CCES Chief Scientist John Drake (ORNL) and the leader of the C-LAMP project, Forrest Hoffman (ORNL). As noted above, a portal has been established at ORNL for the C-LAMP project, and it is expected to being receiving data in April 2007.

A.5.4 Open Science Grid (OSG, SciDAC Physics SAP) and the Center for Enabling Distributed *Petascale Science (CEDPS, SciDAC CS CET)*

The three-way group of OSG, CEDPS, and ESG has been working together to define areas of interaction and complementary work. Don Middleton (NCAR) met with Thomas Ndousse and colleagues from OSG and CEDPS in a March half-day meeting at the San Diego Supercomputing Center (SDSC) in order to advance this activity.

A.5.5 Visualization and Analytics Center for Enabling Technologies (VACET, SciDAC Visualization CET)

LANL team member (Phil Jones) and LLNL team members (Bob Drach, and Dean Williams) participated in the VACET project-wide meeting held in Utah. Science projects supported by ESG and the SciDAC awarded climate application were discussed and plans for collaborations are in progress. Integration of the VACET product visTrails and an ESG supported client application is now under way. VisTrails utilizes a visualization pipeline architecture that provides infrastructure to combine existing client applications into one environment. A particularly powerful use of visTrails is parameter exploration.

A.5.6 Workshop (DOE Cybersecurity R&D Challenges for Open Science: Developing a Roadmap and Vision)

Representing ANL, ESG, and NCAR, and the science/data community in general, Middleton & Ananthakrishnan attended the DOE-sponsored Cybersecurity Workshop, which was held in January 2007 at AGU headquarters in Washington, D.C. Middleton provided a short position paper along with concerns and scenarios. Rachana presented the whitepaper "Trust-Root Provisioning and Validation Facilities" by Frank Siebenlist, Michael Helm, Rachana Ananthakrishnan, and Ian Foster, which discusses deployment concerns for DOE projects like ESG.

A.5.7 TeraGrid (NSF)

Don Middleton continued to participate in TeraGrid meetings and activities, including the extension of ESG to function as a science gateway utilizing TG resources. Middleton attended the January 2007 TG Data Workshop in San Diego, CA.

A.5.8 Community Access to Global Cloud Resolving Model Data And Analyses (SciDAC Climate SAP)

Don Middleton (NCAR) attended the NSF CMMAP (Center for Multi-scale Modeling of Atmospheric Processes) in Kauai, Hawaii, meeting with colleagues Karen Schuchardt (PNNL) and John Helley (SDSC). This group worked on plans to develop an overall program plan that was complimentary relative to our specific project goals across SciDAC-2 and NSF Center for Enabling Technology goals.

A.5.9 A Data Domain to Model Domain Conversion Package (DMCP) for Sparse Climate Related Process Measurements (SciDAC Climate SAP)

ANL and ORNL team members (Dan Fraser and David Bernholdt) met with members of the DMCP project in late February and early March to explore possibilities for collaboration. Since much of what was discussed would take us beyond the current scope and funding of ESG-CET, we are currently deciding how to proceed.

A.5.10 Scientific Data Management Center for Enabling Technology (SciDAC CS CET)

The experience with the DataMover-Lite (DML) client component in ESG, led us to consider its use in the SDM center for moving files from sites that have one-time-password (OTP) security or other highly secure systems. Rather than "pulling" files into the client's site as done in ESG, the SDM center will use a DML for "pushing" files out of highly secure sites to their destinations. The SDM center is planning to use this technology for a combustion project as well as for part of the workflow in a fusion project.

A.5.11 North American Regional Climate Change Assessment Program (NARCCAP, multi-agency)

Don Middleton (NCAR) continued to serve in the co-PI role for this project which will host its data products via the ESG system. During this reporting period, we refined the project's data management plan and began to develop the required software and system infrastructure.

Appendix B Outreach, Papers, Presentations, Posters, and Portals

Outreach activities, papers, talks, and posters presented during the following time periods:

A.6 Outreach Activities:

A.6.1 October 1, 2008 through March 31, 2009

- *Advanced Scientific Computing Research (ASCR) Networking Requirements Workshop, Washington D.C., April 15-16, 2009*

  Dean N. Williams represented the Earth System Grid Center for Enabling Technologies (ESG-CET), where ESG-CET was one of the case studies examined at the workshop. The workshop goal was to accurately characterize the networking requirements of science funded by the ASCR Program Office.

- *Visit to CMCC*

  Ian Foster visited the Italian Euro-Mediterranean Center for Climate Change (CMCC) to participate in their scientific advisory board. He discussed opportunities for collaboration in distributed data management.

- *Institute Simon Pierre Laplace, Paris, France, March 27, 2009 and ORAP Workshop, Lille, France, March 26, 2009*

  Don Middleton gave an invited presentation at the bi-annual ORAP workshop; a France-centered activity aimed at advocacy in the high-performance and large data area, and with goals of becoming a more general E.U. initiative. Middleton's presentation was entitled *"Building a Global Data Federation for Climate Change Science: The Earth System Grid and International Partners"*. After this event in Lille, France, Middleton traveled to Paris and gave the presentation again to a climate modeling and IT group with the Institute Simon Pierre Laplace (IPSL) at the University of Paris Jussieu campus. This meeting also included discussion of the upcoming CMIP-5 process, and how IPSL would collaborate with ESG to function as a data provider.

- *Open Science Grid Workshop, Baton Rouge, Louisiana, March 3-5, 2009*

  Don Middleton gave an invited presentation at the OSG Workshop entitled *"The Earth System Grid and Global Federation of Climate Model Data"*. The presentation combined an overview of the science domain (OSG participants were mostly in the physics area), ESG's goals, and potential for collaboration with OSG in software delivery.

- *DAARWG Workshop in Saint Louis, Mississippi, February 3-5, 2009*

  Dean N. Williams is a member of the NOAA Advisory's Board Data Archive and Access Requirements Working Group (DAARWG), an important working group of NOAA's Science Advisory Board (SAB). DAARWG evaluates data archiving and access requirements from all of NOAA's observing systems and computational models, as well as from relevant non-NOAA sources. Its charter is to provide scientific advice and broad direction regarding the wide range of data, information, and products that NOAA should archive, and ways in which this agency can best provide access to them.

- *NASA Goddard Institute for Space Studies (GISS) Workshop on Advancing the Scientific Value of the CMIP/IPCC Process, New York City, NY, November 17-19, 2009*

  Don Middleton attended this meeting representing ESG and other projects, and led a discussion session on data management and access, including a briefing to the attendees on ESG-CET.

A.6.2 April 1, 2008 through September 30, 2008

- *Cyberinfrastructure Summer Institute for Geoscientists (CSIG), San Diego, CA., August 11-15, 2008*

  Don Middleton gave an invited presentation entitled *"Cyberinfrastructure for Global Earth System Data Federation".* The presentation had a primary focus on ESG, along with additional information on related projects and data analysis and visualization capabilities.

- *World Meteorological Organization (WMO) Information System Intercommission Coordination Group (ICG-WIS), Brazilia, Brazil, July 14-17, 2008*

  Don Middleton serves as an expert advisor to this high-level coordination group for the development of next generation globally federated data systems for climate, weather, hydrology, oceanography, and other U.N./WMO program elements. Middleton attended a meeting of this group in August in Brazilia, Brazil and verbally briefed the council on ESG-CET activities and implications related to the next IPCC process and related federated data systems.

- *DAARWG Workshop in Ashville, NC*

  Dean Williams accepted the invitation to join the NOAA Advisory's Board Data Archive and Access Requirements Working Group (DAARWG), an important working group of NOAA's Science Advisory Board (SAB). DAARWG evaluates data archiving and access requirements from all of NOAA's observing systems and computational models, as well as from relevant non-NOAA sources. Its charter is to provide scientific advice and broad direction regarding the wide range of data, information, and products that NOAA should archive, and ways in which this agency can best provide access to them.

- *Global Earth Observing System of Systems (GEOSS)*

  GEOSS is a primary thrust of the Group for Earth Observations (GEO), which is a United Nations program coordinated under the auspices of the World Meteorological Organization (WMO). Don Middleton attended several days of GEOSS meetings in Boulder, CO. in September 2008. Middleton also gave an invited presentation entitled "Global Federated Climate and Weather Data Systems" and focused on ESG-CET in the area of climate and the THORPEX Interactive Grand Global Ensemble (TIGGE) in the area of weather.

- *Steering Committee for HPDG08*

  Peter Fox is a member of the Steering and Programme Committee for HPDG08 (High Performance Data Grid 2008) to be held in Dunedin, New Zealand Dec 3-4, 2008.

- *EGU Session Convener*

  Peter Fox convened a European Geophysical Union meeting session on "The Grid for Geosciences and Geoscience Applications for the Grid". Fox was co-convener of the Geosciences applications on Grid session also at EGU 2009.

- *ESIn Editor*

  Peter Fox is an *Earth Science Informatics* (ESIn) editor for a special issue on "Grid computing for Geosciences, Earth Science Informatics".

- *World Meteorological Organization Expert Team on GISC's and DCPC's Meeting, Darmstadt Germany, June 6-12, 2008*

  Don Middleton serves on this WMO expert team, and gave a presentation at the annual meeting that included a status update on ESG.

- *Ocean Observatory Initiative*

  Ian Foster accepted an invitation to join the Advisory Committee for the NSF Ocean Observatory Initiative. He attended his first meeting in Dallas Texas and was able to discuss opportunities for joint work with ESG-CET.

A.6.3 October 1, 2007 through March 31, 2008

- *ASCAC-BERAC Charge Committee*

  Ian Foster and Dean Williams served on the Joint Advanced Scientific Computing Advisory Committee (ASCAC) and the Biological and Environmental Research Advisory Committee (BERAC) Subcommittee to investigate barriers and bottlenecks to achieving successful outcomes of complementary investments—specifically in climate modeling by ASCR and BER. Ian and Dean contributed to a 19-page report that was transmitted to Ray Orbach, Director of the DOE Office of Science.

- *Global Organization for Earth System Science Portal (GO-ESSP) Mini-Workshop*

  In mid-October, Steve Hankin, Don Middleton, and Dean N. Williams attended the GO-ESSP mini-workshop which focused on establishing open access to distributed climate data archives.

- *SDM Center All-hands Meeting*

  Hosted by Arie Shoshani, Don Middleton and Dean Williams attended the SDM Center All-hands Meeting where climate analysis and workflow use cases where discussed. Don and Dean represented ESG-CET and made technology collaboration contacts (e.g., for Kepler and VisTrails).

A.6.4 April 1, 2007 through September 30, 2007

- *Presentation: Co-Chair of the GO-ESSP Workshop in Paris, France*

  As Principal Investigators and members of the organizing committee, Dean Williams, Don Middleton, and Steve Hankin attended the 6th Annual Global Organization for Earth System Science Portal (GO-ESSP) Workshop promoting this effort's goals and objectives. The GO-ESSP is a collaboration designed to develop a new generation of software infrastructure that will provide distributed access to observed and simulated data for the climate and weather communities. GO-ESSP will achieve this goal by developing individual software components and by building a federation of frameworks that can work together using standards agreed upon by its participants. The GO-ESSP portal frameworks will provide efficient mechanisms for data discovery, access, and analysis of the data. Participants shared their progress in developing software infrastructure that facilitated discovery, acquisition, and analysis of climate date. Particular interest was expressed on

current and future integration activities that facilitate community analysis of widely distributed climate data archives (e.g., CMIP3 (IPCC AR4) and CMIP4 (IPCC AR5)).

- *SciDAC 2007 Organizing Committee*

  Ian Foster and Dean Williams served on the SciDAC 2007 organizing committee, which selected topics that represent state-of-the-art for a given scientific area and suggested appropriate speakers on each topic. Ian was the committee organizer for "Grids/Networking", and Dean served as both the committee organizer for the "Climate Community" and as a "Session Chair" at the conference. The OC also suggested topics and presenters for invited poster sessions. For each topic area, the respective OC member was responsible for peer-review presenter abstracts before the conference, and of proceedings papers immediately after the conference.

A.6.5 October 1, 2006 through March 31, 2007

- *WMO Information System (WIS)*

  Don Middleton (NCAR) serves on multiple WMO WIS committees, and provided updates and overviews on ESG-CET at an intercommission coordination group in Beijing, China, and again at the international "Extraordinary Assembly" of WMO, a technical conference for information technology, held in Seoul, Korea November 2007.

A.7 Papers:

A.7.1 October 1, 2008 through March 31, 2009

- *Earth System Modelling (ESM) Software, Tools, and Environments Book*

  Robert Drach, Steve Hankin, Don Middleton, and Dean N. Williams, provided written input for chapter 5 (i.e., entitled, "*IO and Post-processing*") of the book titled, "*Earth System Modelling (ESM) Software, Tools, and Environments*". They authored the sub-sections: "*Data Representation*" and "*Data Analysis and Visualization*". More written input highlighting "*The Earth System Grid: Distributed and Uniform Access to ESM Data*" will be provide for chapter 7 (i.e., entitled, "ESM-*Data Archives in times of the GRID*").

- *SciDAC Review Article*

  D. N. Williams, R. Ananthakrishnan, D. E. Bernholdt, S. Bharathi, D. Brown, M. Chen, A. L. Chervenak, L. Cinquini, R. Drach, I. T. Foster, P. Fox, S. Hankin, V. E. Henson, P. Jones, D. E. Middleton, J. Schwidder, R. Schweitzer, R. Schuler, A Shoshani, F. Siebenlist, A. Sim, W. G. Strand, N. Wilhelmi, M. Su. "*The Planet at Their Fingertips: Climate Modeling Data Heats Up*", Spring 2009.

  The ESG-CET team completed the SciDAC Review Article entitled, "The Planet at Their Fingertips: Climate Modeling Data Heats Up". The article talks about the increasing importance of climate modeling and the tremendous need for the Earth System Grid to allow fast and accurate access to hundreds of petabytes.

  (URL: http://www.scidacreview.org/0902/html/esg.html)

- *Paper in the Bulletin of the American Meteorological Society (BAMS)*

  D N Williams, R Ananthakrishnan, D E Bernholdt, S Bharathi, D Brown, M Chen, A L Chervenak, L Cinquini, R Drach, I T Foster, P Fox, D Fraser, J Garcia, S Hankin, P Jones, D E Middleton, J

Schwidder, R Schweitzer, R Schuler, A Shoshani, F Siebenlist, A Sim, W G Strand, M Su, N. Wilhelmi, "*The Earth System Grid: Enabling Access to Multi-Model Climate Simulation Data*", in the Bulletin of the American Meteorological Society, February 2009.

(URL: http://ams.allenpress.com/perlserv/?request=get-abstract&doi=10.1175/2008BAMS2459.1)

This article, by the ESG-CET team, follows the Meehl et al., 2007 *BAMS* article on *"The WCRP CMIP3 Multi-Model Dataset: A New Era in Climate Change Research"*.

- *Scientific Grand Challenges: Challenges in Climate Change Science and The Role of Computing at the Extreme Scale Report*

  Dean N. Williams, Don Middleton, et al. co-authored the section, "Data, Visualization, and Computing Productivity" in the Scientific Grand Challenges: Challenges in Climate Change Science and The Role of Computing at the Extreme Scale Report, November 6-7, 2008.

- *Developing Service-Oriented Applications in a Grid*

  Garcia, J., Fox, P., West, P. and Zednik, S. 2009, Developing Service-Oriented Applications in a Grid Environment. Experiences Using the OPeNDAP Back-End-Server Presentations, Earth Science Informatics, in press.

A.7.2 April 1, 2008 through September 30, 2008

- *Earth Science Informatics Article*

  Developing Service-Oriented Applications in a Grid Environment: Experiences Using the OPeNDAP Back-End-Server, 2008, J. Garcia, P. Fox, P. West. S. Zednik.

  Submitted to Earth Science Informatics as part of the special issue on "Grid Computing for Geosciences".

- *Federation*

  Robert Schuler and Ann Chervenak, wrote a federation document for the ESG-CET team entitled, "Search Metadata: Storage and Sharing Considerations". In addition, they gave a talk on "Federated Metadata" at the ESG All-Hands Meeting in April 2008.

- *SciDAC Review Highlights*

  The ESG-CET team prepared an article for the SciDAC Review highlighting ESG-CET. Several of the SciDAC projects that reported at the SciDAC 2008 conference were asked to present a one-page summary of their project, together with an image and a caption.

A.7.3 October 1, 2007 through March 31, 2008

- *Paper in the Cyberinfrastructure Technology Watch (CTWatch)*

  D. N. Williams, D. E. Bernholdt, I. T. Foster, and D. E. Middleton, 2007: *The Earth System Grid Center for Enabling Technologies: Enabling community access to petascale climate datasets*. *Cyberinfrastructure Technology Watch (CTWatch) Quarterly*, Vol. 3, number 4.

- *International Symposium on Collaborative Technologies and Systems*

  W. Kendall, M. Glatter, J. Huang, F. Hoffman, and D. E. Bernholdt, 2008: Web enabled collaborative climate visualization in the Earth System Grid. Invited paper for the 2008 International Symposium on Collaborative Technologies and Systems (CTS 2008), to appear.

- *LLNL 2007 Computation Directorate Annual Report*

  A.A. Mirin, and D.N. Williams, 2008: Climate modeling – Conducting research and providing enabling technology. *LLNL 2007 Computation Directorate Annual Report*.

A.7.4 April 1, 2007 through September 30, 2007

- *Poster and Paper: SciDAC '07 Conference*

  The ESG team presented a peer-reviewed paper in the SciDAC 2007conference proceedings. The complete citation is: R Ananthakrishnan, D E Bernholdt, S Bharathi, D Brown, M Chen, A L Chervenak, L Cinquini, R Drach, I T Foster, P Fox, D Fraser, K Halliday, S Hankin, P Jones, C Kesselman, D E Middleton, J Schwidder, R Schweitzer, R Schuler, A Shoshani, F Siebenlist, A Sim, W G Strand, N. Wilhelmi, M Su, and D N Williams, "Building a Global Federation System for Climate Change Research: The Earth System Grid Center for Enabling Technologies (ESG-CET)", in the Journal of Physics: Conference Series, SciDAC '07 conference proceedings.

A.7.5 October 1, 2006 through March 31, 2007

- *Paper and Presentation: eScience'06*

  Ann Chervenak (ISI/USC) presented a peer-reviewed paper at eScience 2006 in Amsterdam, Netherlands. The complete citation is: Ann Chervenak, Jennifer M. Schopf, Laura Pearlman, Mei-Hui Su, Shishir Bharathi, Luca Cinquini, Mike D'Arcy, Neill Miller, and David Bernholdt, Monitoring the Earth System Grid with MDS4, in *Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*, page 69, Los Alamitos, CA, USA, 2006, IEEE Computer Society.

- *Position Paper: DOE Cybersecurity Workshop*

  Representing ESG, NCAR, and the science/data community in general, Don Middleton (NCAR) presented a short position paper at the DOE-sponsored Cybersecurity Workshop, which was held in January 2007 at AGU headquarters in Washington, D.C.

A.8 Talks:

A.8.1 October 1, 2008 through March 31, 2009

- *Collaboratories Workshop*

  Dean N. Williams et al., presented "Composing Collaboratories: Earth System Grid Center for Enabling Technologies (ESG-CET)" at the Collaboratories Workshop held in Chicago, IL, February 24-25, 2009.

- *Osni Marques Visit to LLNL*

  Dean N. Williams presented "The Earth System Grid Center for Enabling Technologies at LLNL" to Osni Marques, ASCR's Computer Science Program Manager for SciDAC. He is working with Dan Hitchcock on all aspects of the Computer Science Program. The talk was given on January 15, 2009.

- *Conference on International Interactive Information and Processing Systems (IIPS)*

  Schweitzer, Roland and K. M. O'Brien, J. Li, A. Manke, J. Malczyk, and S. Hankin, "A General Purpose System for Server-side Analysis of Earth Science Data", 25th Conference on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology Phoenix, AZ, Wednesday, 14 January 2009.

- *MAGIC Teleconference*

  Don Middleton and Dean N. Williams presented the Earth System Grid at the MAGIC Meeting Teleconference. The meeting was held on December 3, 2008.

- *Supercomputing 2008*

  Dean N. Williams gave a presentation on "Climate Change: Current Knowledge and Future Challenges" at the 2008 Supercomputing Conference for High Performance Computing, Network, Storage and Analysis, held in Austin, TX, November 15-21, 2008. Dean presented efforts to build up and disseminate climate data via the Earth System Grid.

- *Extreme Scale Computing Challenges in Climate Change Science Workshop*

  Don Middleton and Dean N. Williams co-chaired the Data, Visualization, and Computing Productivity session at the Extreme Scale Computing Challenges in Climate Change Science Workshop. Other responsibilities at the workshop included: organization, hosting breakout sessions, presentations, and providing written input for the final workshop report. The workshop was held in Washington D.C., November 6-7, 2008.

- *Presentation at GO-ESSP Community Meeting*

  Schweitzer, Roland and K. M. O'Brien, A. Manke, J. Malczyk, and S. Hankin, "*A General Purpose System for Server-side Analysis of Earth Science Dat*a", 6th GO-ESSP Community Meeting, Seattle WA, 19, September 2008.

- *Presentation at EGU 2009*

  The Climate-G testbed: towards a large scale data sharing environment for climate change by G. Aloisio et al.

A.8.2 April 1, 2008 through September 30, 2008

- *Net@Edu Meeting*

  Don Middleton provided the keynote address at the annual Net@Edu meeting in Tempe, Arizona in February 2008 (this was not reported in the previous ESG-CET report for that time period). Entitled *"Cyberinfrastructure and Emerging Scientific Data and Knowledge Systems"*, Middleton's presentation highlighted ESG-CET as a primary focal point of the talk.

- *North American Regional Climate Change Assessment Program (NARCCAP) PI and Community Meeting*

  NARCCAP is a multi-agency (NSF, DOE, NOAA, EPA) regional climate modeling program that is closely affiliated with ESG-CET, and providing data services to the community using ESG cyberinfrastructure. Don Middleton contributed to the organization of this meeting and presented an opening talk entitled *"NARCCAP Data: The Earth System Grid (ESG), Data & Knowledge Systems,*

and a Few Useful Tools". We also held a hand-on training session for community members on using the ESG-based NARCCAP data system, and it was well received and effective.

- *Virtual Climate Change Research Institute Workshop*

  Ann Chervenak gave a presentation on "The Earth System Grid: Turning Climate Datasets into Community Resources" at the Virtual Climate Change Research Institute Workshop held at USC on June 25, 2008. A group of paleoclimatologists is proposing a new NSF Science and Technology Center and hopes to collaborate with ESG-CET.

- *World Meteorological Organization (WMO) ET-WISC Meeting*

  In June 2008, Don Middleton attended a meeting in Darmstadt, Germany of the Expert Team on the World Meteorological Organization (WMO, a program of the United Nations) Information System (WIS) Global Federated Systems (ET-WISC). Middleton provided a presentation on federated scientific data systems included a briefing on ESG-CET for the WMO expert team.

- *Washington trip to meet with Sponsors*

  On June 27th, Dean Williams spent a day in Washington, D.C. with sponsors from the DOE Office of Science SciDAC and from the Office of Biological and Environmental Research (OBER). Dean presented information on ESG-CET and discussed the progress of this project. The meeting also provided an opportunity to introduce Dean and ESG-CET to Susan Turnbull, the new ESG-CET SciDAC Program Manager.

- *ESnet Climate meeting at LBNL with Dan Hitchcook*

  David Bernholdt contributed to a presentation by John Drake (lead PI for the SciDAC CCSM effort) discussing data volumes and networking needs anticipated for CMIP5 (IPCC AR5), presented at the BER-ESnet workshop on 4 August. Dean Williams was also present at this meeting and presented information on ESG-CET.

- *BER Program Manager Wanda Farrell's visit to ORNL*

  On August 6th, BER Program Manager Wanda Farrell visited ORNL. David Bernholdt contributed to a presentation by John Drake on modeling-related activities at ORNL, including ESG-CET.

- *Presented ESG-CET to Christine Chalk*

  On August 6th, BER Program Manger Christine Chalk visited LLNL. Dean Williams presented ESG-CET and discussions took place covering the use of ESG-CET for SciDAC's Biology & Environment scientific area.

- *Presented ESG-CET to David Skinner and Horst Simon*

  On August 13th, Dean Williams visited NERSC to discuss the ESG-CET and the possible need for more storage space on NERSC's long-term storage facility. In addition, future conversations were planned on NERSC as an ESG "Data Node" and as a compute server.

- *BER Program Manager Anjuli Bamzai visit to ORNL*

  On September 11th, BER Program Manager Anjuli Bamzai visited ORNL, where David Bernholdt gave a presentation on ESG.

- *Global Organization for Earth System Science Portal (GO-ESSP) Workshop*

  At the GO-ESSP workshop, the following presentations were presented by ESG-CET team members: Luca Cinquini - "The ESG-CET Gateway: A Distributed and Federated Architecture for Data Search, Access and Analysis", Nate Wilhelmi - "ESG-CET Demonstration", Eric Nienhouse - "The Challenges of Building Rich Web Clients Geo-Spatial Applications – Lessons Learned", and Roland Schweitzer - "Server-side OPeNDAP Analysis – Concrete steps toward a generalized framework via a reference implementation using F-TDS".

- *Presentation at EGU 2008, IN09, The Grid for Geosciences*

  Grid-enabled OPeNDAP Hyrax servers: parallel/multiple back-end services and remote netCDF invocation (RNI), by J. Garcia, S. Zednik, P. West and P. Fox.

A.8.3 April 1, 2007 through September 30, 2007

- *Presentation: Co-Chair of the IPCC WG1*

  Dean Williams and Robert Drach demonstrated ESG-CET to Dr. Susan Solomn prior to her April 2007 LLNL "Director Distinguished Lecturer" series presentation on the scientific findings of the IPCC Working Group I (WG1), which were recently published in its fourth comprehensive assessment report (AR4). Dr. Solomon is a senior scientist at the Aeronomy Laboratory (a National Oceanic and Atmospheric Administration facility) and has served as co-chair of the IPCC Working Group I (WG1).

- *Presentation: Fusion Energy Science Community -- Dr. William Tang*

  Dean Williams (LLNL) gave a presentation on ESG-CET to Dr. William Tang, the Chief Scientist at the Princeton Plasma Physics Laboratory (PPPL), a national laboratory for fusion research. Dr. Tang played a prominent leadership role for the Department of Energy's development multi-disciplinary program in advanced computational science, (i.e., the Scientific Discovery through Advanced Computing (SciDAC)). We discussed ways in which ESG-CET might be used to assist the DOE's Fusion Energy science community. This collaboration also involves the use of LLNL's computing resources, such as the Green Data Oasis and the Green Linux Capacity Cluster (GLCC).

- *PCMDI Program Review:*

  Dean Williams presented a PowerPoint presentation on ESG-CET, subtitled: "Data and Software: Turning Climate Datasets into Community Resources" to the PCMDI Program Review Committee on August 27, 2007 in Livermore, CA.

- *Poster and Presentation: Climate Change Prediction Program (CCPP) '07 Conference*

  Representing ESG, Dave Bernholdt and Dean Williams presented the ESG-CET poster at the September 2007 Climate Change Prediction Program (CCPP) conference, which was held in Indianapolis, Indiana. The poster was entitled: "Building a Global Infrastructure for Climate Change Research". Dean also presented a PowerPoint presentation on ESG-CET, entitled: "Data and Software Infrastructure for the Global Climate Community".

- *Presentation: World Meteorological Organization Information System (WMO-WIS) Intercommission Coordination Group*

  The World Meteorological Organization (WMO) is in the process of designing and building its next generation global information system, an effort known as WMO-WIS. While WMO has long had an

operational network for meteorological observations and warnings, the new system is to provide data management and access across the various WMO directorates, thus encompassing weather, climate, oceans, and more. Don Middleton serves on the Expert Team chartered with architecting and designing the federation of national and international systems and also serves as an advisor for the high-level Intercommission Coordination Group (ICG-WIS). Middleton gave a presentation at the group's recent September meeting in Reading, U.K. that included an update on ESG-CET, and outlined opportunities for collaboration and idea exchange in the areas of metadata, federation, and virtual organizations.

A.8.4 October 1, 2006 through March 31, 2007

- *Presentation: VACET All-hands Meeting*

  Bob Drach (LLNL) gave an ESG-CET presentation at the VACET All-hands meeting held at the University of Utah in Salt Lake City, Utah.

- *Presentation: Boulder Data Group*

  Don Middleton (NCAR) presented a talk entitled, "The Earth System Grid: Delivering Petascale Climate Data to a Global Community" at the inaugural meeting of the new Boulder Data Group, held in Boulder, CO.

- *Presentation: SC2006*

  ESG-CET activities were featured at booths sponsored by ANL, Global Grid Forum, LBNL, NCAR, and ORNL in the Exhibit Hall of SC2006, Tampa Bay, FL.

- *Presentation: DOE LLNL Program Review*

  Dean Williams (LLNL) presented ESG-CET at the DOE PCMDI Program Review held in November 2007 at LLNL, CA.

- *Presentation: Curator Workshop*

  Dean Williams (LLNL) presented ESG-CET information at the Curator Workshop on Metadata held at GFDL in Princeton, NJ.

A.9 Posters:

A.9.1 October 1, 2008 through March 31, 2009

- *Climate Change Prediction Program (CCPP) Meeting*

  Dean N. Williams, et al., presented the poster titled, "The Earth System Grid: Scientific Discovery Made Easy", at the 2009 CCPP meeting in Washington, D.C., 6-9 April 2009.

- *American Geophysical Union (AGU) Fall Meeting*

  Dean N. Williams, et al., presented the poster titled, "The Earth System Grid: Scientific Discovery Made Easy" at the AGU Fall Meeting in San Francisco, 15-19 December 2008.

- *Global Ocean Data Assimilation Experiment Symposium*

  Hankin Steven, et al., "Climate and Forecast (CF) Conventions for NetCDF – The Foundation of GODAE Data Interoperability", (poster) Global Ocean Data Assimilation Experiment final symposium -- Observing and Forecasting the Ocean, Nice, France, 10-15 November 2008.

A.9.2 April 1, 2008 through September 30, 2008

- *SciDAC '08 PI Meeting in Denver, CO*
  David Bernholdt, Ian Foster, Don Middleton, and Dean Williams (the ESG-CET executive team) attended the SciDAC PI meeting held in Denver, CO, where Don and Dean presented a poster on ESG-CET. Besides providing a great opportunity for the executive to intermingle with other Advanced Scientific Computing Research (ASCR) computer science efforts, this meeting provided an opportunity to discuss topics in areas that may assist future ESG development, including performance tools, application development frameworks, data management/analytics, and visualization.
- *European Geophysical Union Meeting*

  In April of 2008, Don Middleton attended the annual European Geophysical Union (EGU) meeting in Vienna, Austria. Middleton presented a poster on ESG-CET, which was developed in concert with the ESG-CET Executive team.

A.9.3 April 1, 2007 through September 30, 2007

- *Poster and Paper: SciDAC '07 Conference*

  Don Middleton presented a poster on ESG-CET at the SciDAC '07 conference held in Boston, MA. Also representing ESG at the conference were Ian Foster, Dave Bernholdt, and Dean Williams. (Taking advantage of the conference, The ESG executive committee held many face-to-face meetings.)

A.9.4 October 1, 2006 through March 31, 2007

- *Poster: SciDAC'07 PI Workshop*

  Dean Williams (LLNL) and Don Middleton (NCAR) presented a poster on ESG-CET at the SciDAC '07 Workshop held in Atlanta, GA.
- *Poster: 3rd WGNE Workshop on Systematic Errors in Climate and NWP Models*

  Dean Williams (LLNL) presented a poster on ESG-CET at the 3rd WGNE Workshop on Systematic Errors in Climate and NWP Models held in February 2007 in San Francisco, CA.

A.10 Portals and Links:

- *CCSM Portal:* Operational portal for the CCSM, CSIM, CLM, NARCCAP, PCM, and POP.
  - http://www.earthsystemgrid.org/
- *WCRP CMIP3 Multi-Model Portal:* Operational portal for the CMIP3 (IPCC AR4) and CFMIP.
  - https://esg.llnl.gov:8443/
- *C-LAMP Portal:* Operational portal for the Carbon Land Model Intercomparison Project (C-LAMP).
  - https://esg2.ornl.gov:8443/